“Rome Wasn’t Digitized in a Day”: Building a Cyberinfrastructure for Digital Classicists
Draft Version 1.3—11/18/10

Alison Babeu
Perseus Digital Library

Introduction

"*You would think that classics is classics, but, in fact, the sensibilities inside the different subdisciplines can be radically different.*" (A scholar as cited in Harley et al. 2010, pg. 72)

"It has always been the case that scholars need to cite primary and secondary texts in retraceable form and argue cogently and replicably from the data to the conclusions, just as it has long been the case that academic language, jargon, abbreviations, and conventions ought to be standardised within (if not between) disciplines. None of the philosophies and practices of the Digital Classics community need therefore be seen as new or unfamiliar." (Bodard 2008)

As these quotes indicate, classics is a complicated and interdisciplinary field with a wide ranging group of sub-disciplines and a seemingly endless variety of technical challenges. Digital classics, for the purpose of this report, is broadly defined as the use of digital technologies in any field related to the study of antiquity. This report seeks to explore the state-of-the-art in digital classics in terms of what projects exist and how they are used, to examine what type of infrastructure already exists to support digital classics as a discipline, and to investigate what types of larger humanities cyberinfrastructure projects already exist and what tools or services have been built that might be repurposed.

This report opens with a brief overview of the history of computing in classics in order to establish important early themes and to set the context for the study of digital classics. This will be followed by a brief look at large multidisciplinary classical digital libraries (Ancient Near East, Greek, Roman) and the types of advanced services they use and are likely to require. A summary overview will then be given of the various disciplines of digital classics, the major projects in each and the major technologies in use. Next, a brief overview of digital humanities user studies will attempt to get at the needs of users of these projects, as no user studies of digital classicists could be found. Afterwards an overview of requirements for a cyberinfrastructure for digital classics will be reviewed along with a survey of relevant projects. The report will conclude with an examination of larger recommendations for a humanities cyberinfrastructure and relevant national and international cyberinfrastructure projects.

Classics and Computers: A Long History

The field of classical studies is an extremely broad one that includes a variety of related disciplines including: ancient history, archaeology, epigraphy, numismatics, papyrology, philology and prosopography and the impact of computing has varied greatly among them.[1] Nonetheless, classical studies and advanced computing technologies have a long history. A useful definition of the field was offered by Lorna Hardwick:

As an academic discipline it is broad in terms of chronology, in geographical provenance and in the range of specialisms on which it relies. It involves the study of the languages, literatures, histories, ideas, religion, science and technology, art, architecture and all aspects of the material and intellectual cultures of the peoples living in and around the Aegean and Mediterranean from the age of Mycenae (c. 1400–1200 BCE) until roughly the seventh century CE (Hardwick 2000).

Melissa Terras also recently offered another helpful summary definition:

Often understood as 'one who advocates the school study of the Latin and Greek classics', this definition belies the complex range of sources and associated research techniques often used by academic Classicists. Varied archaeological, epigraphic, documentary, linguistic, forensic and art historical evidence can be consulted in the course of everyday research into history, linguistics, philology, literature, ethnography, anthropology, art, architecture, science, mythology, religion and beyond. Classicists, have by nature and necessity, always been working across disciplinary boundaries in a data-intensive research area, being 'interdisciplinary, rather than simply un-disciplined'. The addition of advanced digital and computational tools to many a Classicists' arsenal of skills should therefore not really come as a surprise, given the efficiencies they afford in the searching, retrieval, classification, labeling, ordering, display and visualization of data (Terras 2010, pg. 172)

Classical studies is thus inherently a very interdisciplinary field and one that has long made use of advanced technology, a theme that will run throughout this paper. A variety of studies have considered the impact of

[1] While each disciplinary subsection will try to provide an overview of the important projects and issues for each discipline, a quick perusal of the table of contents (http://www.worldcat.org/isbn/0754677737) of the just published *Digital Research in the Study of Classical Antiquity* (Bodard and Mahony 2010) illustrates the diversity of research in digital classics.

computing on classical studies and this section will provide a brief overview of them in order to set the context of more recent developments in the field of "digital classics."

The evidence that classicists draw on can range from buildings, inscriptions, artifacts, art objects, and written evidence such as poetry, drama, narrative histories, and philosophical works. One particularly challenging aspect is that many of the textual sources are fragmentary. Hardwick notes that:

> Even where we possess a more or less complete form, texts have generally survived via manuscripts copied and recopied in late antiquity or medieval times. Almost all of the material evidence has to be excavated and/or reconstructed. Even objects which have survived relatively intact often lack their original contexts. Thus, far from consisting of a fixed, closed body of knowledge, as used to be imagined when Classical Studies had the reputation of being the ultimate 'canonical' field of study, the discipline often involves considerable experimentation, conjecture and hypothesis (Hardwick 2000).

Although classics is often considered to be a field of fixed knowledge, the research that will be discussed in this paper will illustrate the inaccuracy of that belief.

To indicate the long history of classics and computing, Stewart et al. (2007) have defined several generations of digital corpora in classics, the first generation simply sought to make texts available online (such as the Latin Library[2]) and relied on community contributions, a second generation of corpora such as the Thesaurus Linguae Graecae (TLG) and Packard Humanities Institute (PHI) Latin library invested in professional data entry and involved scholars in the checking of all the texts, both to correct transcriptional errors and to provide a consistent markup scheme.[3] This generation also involved the development of BetaCode by classicists to capture ancient languages such as Greek and Coptic. A third class of corpora that involved taking professionally entered text and semantically marking it up in SGML/XML such as with the markup designed by the Text Encoding Initiative (TEI)[4] evolved in the 1980s and included the Perseus Digital Library. A fourth generation of corpora involved image-front collections that provided users with page images that included hidden uncorrected OCR that could be searched, and this strategy popularized in the 1990s has driven mass digitization projects such as Google Books[5] and the Open Content Alliance.[6] Stewart et al., however, called for a fifth generation of corpora, that synthesized the strengths of the four previous generations, while also including the following features: they will allow decentralized contributions from users, they will utilize automated methods to create both scalable and semantic markup, and they will "synthesize the scholarly demands of capital intensive, manually constructed collections" such as Perseus, the TLG and the PHI, with "the industrial scale of very large "million book" libraries now emerging."

The early exploration of the potential of classics and computing is demonstrated in an article written by James McDonough in 1959. He opened with a discussion of how it took James Turney Allen almost 43 years to create a concordance of Euripides, a task that a newly available IBM computer could do in 12 hours. McDonough continued with the now canonical example of how Father Roberto Busa used a computer to create a concordance to the works of Thomas Aquinas.[7] These examples were used by McDonough to explain that computers could help revolutionize studies by performing exceptionally time consuming manual tasks such as the creation of concordances, textual emendation, auto abstracting of articles, and most importantly, the collection and collation of manuscripts. Although McDonough observed that "machines now make economically feasible a critical edition in which the exact reading of *every* source could be printed in full", this phenomenon still has yet to occur, a point to which we shall return in our discussions of digital critical editions and manuscripts.

McDonough optimistically predicted that new computing technologies would convince classicists to take on new forms of research that were not previously possible, arguing that classicists were entering: "A new era in

---

[2] http://thelatinlibrary.com/

[3] A special open source tool named Diogenes (http://www.dur.ac.uk/p.j.heslin/Software/Diogenes/) was created by Peter Heslin to work with these two corpora (TLG, PHI), as many scholars had criticized the limited usability as well as searching and browsing features of these two "commercial" databases.

[4] http://www.tei-c.org/index.xml

[5] http://books.google.com

[6] http://www.archive.org

[7] The work of Father Busa is typically considered to be the beginning of classical computing (Crane 2004), as well as literary computing and corpus linguistics (Lüdeling and Zeldes 2007).

scholarship, a golden age in which machines perform the servile secretarial tasks, and so leave the scholar free for his proper function, interpretive scholarly re-search....". He concluded his piece with three recommendations: 1) that all classicists request the machine tape for their editions be given to them; 2) that classical studies associations work together to found and support a center that will record the complete texts of at least all major Latin and Greek authors; 3) that relevant parties should increase their comprehensive bibliographic efforts. As a final thought, McDonough returned to the lifetime work of James Turney Allen. "That such techniques as this article attempts to sketch were not available to Professor Allen at the turn of the century is tragic," McDonough offered, "If they be not extensively employed from this day forth by all interested in scholarship, it will indeed by a harsh commentary on our intelligence" (McDonough 1959). McDonough's points still ring true today, about the importance of making all primary data such as manuscripts and texts available, the need for classical associations to work together, and the need for scholars to give up obsessing over "slavish" tasks and return to the more important work of humanistic interpretation of sources.

Over 30 years later in 1991, J.D. Bolter offered a detailed analysis of how one particular feature of the Internet, hypertext, offered great new potential for recording and presenting scholarship (Bolter 1991). Bolter contended that studies for the last two centuries had been defined by the qualities of the printed book where the main goal of scholarship had been to "to fix each text of each ancient author: to determine the authenticity of works ascribed to an author and for each work to establish the *Urtext*—what the author actually wrote, letter for letter." This is the essential work of a classical philologist or creator of a critical edition, to study all the extant manuscripts of an author, to "reconstruct" a text, and then to list all the text variants (or at least the most important) found in the original sources with explanations in an apparatus criticus. While postmodern literary theory had challenged the ideal of the Urtext in many disciplines, Bolter submitted, classical studies had remained largely unaffected. Nonetheless, Bolter believed that the nature of hypertext was affecting how classicists perceived the nature of their texts. "For hypertext now challenges the Urtext not in the jargon of postmodern theory," Bolter explained, "but practically and visibly in the way that it handles text." Hypertext, he posited, might lead scholars to focus less on establishing an exact Urtext and instead start exploring the *connections* between texts and instead "emphasize the continuity between the ancient text and its ancient, medieval, and modern interpretations." The need for digital editions to reflect a more sophisticated tradition of textual transmission and textual variation is a theme that is echoed throughout the literature of digital classics.

An even more extensive exploration of how new technologies might affect the discipline of classics was provided by Karen Ruhleder in 1995, albeit with an exclusive focus on the TLG. One challenge Ruhleder underscored was that "humanists themselves have been more interested in applying computing technologies to their work in detailed studies of the impact of computing technologies on their disciplines" a criticism echoed by Christine Borgman in her recent examination of the digital humanities (Borgman 2009). Ruhleder surveyed how using the TLG had affected the daily work of classicists, how it had changed their relationship to the textual materials they used, and how it affected both social relations among classicists and their disciplinary infrastructures. She conducted sixty unstructured interviews with classicists and concentrated on work in literary scholarship and textual criticism. Ruhleder observed that the work of classical scholarship was often like detective work and that scholars' questions typically included "manuscript authorship and authenticity, social relationships between different groups or classes in ancient Greek society, and the different meanings of a word or phrase over time." Classical scholars used analytical techniques to weigh and interpret evidence and utilized tools "to locate particular pieces of evidence within texts." As Bolter explained previously, the nature of textual evidence for classicists is complicated because materials are often fragmentary or questionable, their transmission is disputable, and the "reconstruction of the original, or urtext, is an important primary activity within classical scholarship."

While the TLG did offer amazing new searching opportunities as well as both breadth and depth of material, Ruhleder criticized the unexamined use of it by classicists. The TLG uses one "best edition" chosen by a special committee of the American Philological Association (APA) for each Greek author and includes no commentaries or apparatus critica, a practice challenged by Ruhleder:

The *TLG* has not only altered the form (book to electronic medium), but it has changed the content and the organization of materials presented in the package. It includes neither critical notes nor other elements of an apparatus criticus and includes only a single edition of each text. These limitations have led to serious criticism, particularly where there is dispute over the version used by the *TLG* (Ruhleder et al. 1995).

Ruhleder also noted that while the corpus may have been broadened in one sense it is also far shallower as critical information has been "decoupled" from the texts. Similar criticism of the TLG has also been offered more recently by Notis Toufexis:

In the absence of detailed contextualization information accompanying the online version of each text, the user who wishes to check the reliability of a given edition (if, for instance, it uses all extant manuscripts of a text or not) has to refer to the printed edition or other handbooks. The same applies to any attempt to put search results obtained by the TLG within the wider context of a literary genre or a historical period. The TLG assumes in a sense that its users have a broad knowledge of Greek literature and language of all historical periods and are capable of contextualizing each search result on their own (Toufexis 2010, pg. 110).

Many classicists that Ruhleder interviewed were also concerned with the authority that was afforded to texts in the TLG, due to their electronic nature. Ruhleder hypothesized that the TLG had affected the work of classicists in three major ways: 1) in terms of the beliefs and expectations classicists had of the materials they worked with 2) the nature of skill sets and expertise and 3) disciplinary infrastructure. In terms of beliefs regarding materials, classicists had previously assumed gaining familiarity with a corpus was a life's work and only happened through constant reading and rereading of the text and that adding to that corpus was a collaborative act.

Ease of searching and finding texts in the TLG, Ruhleder proposed, now left scholars free to pursue other work such as scholarly tool building or the creation of electronic texts, but this process was not without its problems:

Of course, tool building is another form of scholarly work in itself, and databanks and electronic texts are a form of "scholarly production." However, this kind of activity has traditionally ranked low; developing an index or concordance ranks above developing teaching materials, but below articles, books, commentaries, and producing new textual editions. Developing computer-based tools is not even on the list (Ruhleder et al. 1995).

The challenge of evaluating scholarly work in digital classics and indeed all of digital humanities as well as the unwillingness of many traditional tenure evaluations to consider digital scholarship are themes that will be seen through the literature.

The second major change identified by Ruhleder, that of shifting skill sets and expertise, largely considered how technical expertise was increasingly being substituted for experience gained over time as well as how classicists increasingly needed more sophisticated technical knowledge to understand the limitations of tools such as the TLG. The third major change, that of challenges to disciplinary infrastructure, Ruhleder used to briefly explore issues that are now major discussion points, the challenges of electronic publication to the traditional print infrastructure and the new level of technical infrastructure and support required to create, disseminate and preserve electronic texts.

After this thorough exploration of the TLG, Ruhleder concluded her piece with a number of larger issues regarding the impact of computing on classics. To begin with, Ruhleder contended that the ways in which the TLG has "flattened" and in some ways misrepresented the corpus of Greek negatively affects the amount of information available to scholars. In addition, as the TLG moved scholars yet one more step away from the original source text, Ruhleder criticized scholars for simply accepting other scholars readings of a text and not ever returning to the primary text to draw their own conclusions. This valid criticism has inspired projects such as Demos,[8] an electronic publication of the Stoa consortium on Athenian democracy, where every statement is linked back to the primary textual evidence (making particular use of the Perseus Project) on which it is based. Ruhleder wondered why more scholars didn't challenge both the nature of the TLG corpus and explore larger questions of how computing was affecting their discipline:

Fundamental paradigmatic questions, methodological discussions, and mechanisms for resource allocation and infrastructural development that are appropriately discussed at the level of the community-of-practice often masquerade as individual problems of skill or resources. "We need to rethink what it means for our discipline to take on a technological character" is reduced to "I feel detached from the text" (Ruhleder et al. 1995).

[8] http://www.stoa.org/projects/demos/home

She hoped that larger discussions would begin to take place in the future and concluded that the main problem was not so much in what systems could or couldn't do but in their users unwillingness to explore their limitations and set them in a broader disciplinary context. Many of the challenges raised by Ruhleder will be explored further in our discussion of the various disciplines of digital classics.

By the early part of this century, many explorations of classics and computers had turned from individual considerations of particular databases to an examination of the Internet. In 2000, Hardwick argued that the impact of the Internet on classics had been largest in the areas of communication, publication, dissemination of research and the development of specialist research tools. She listed a number of importance advances, including: improved access to primary texts and secondary research, rapid search tools, research databases, rapidly updated specialist bibliographies, electronic journals, more quickly reviewed academic publications, and new potential with electronic discussion lists and conferences. Teaching in classical studies had also changed over the last thirty years, Hardwick argued, with an increasing focus on history and culture and a lower focus on language based learning. Though Hardwick also noted that whereas students used to enter college with a fair amount of linguistic learning, they were now first encountering Greek and Roman culture and history in a variety of venues, which had consequently reinvigorated their interest in language learning. Another interesting insight offered by Hardwick was that the growing availability of data online was "blurring the lines" between teaching and research, helping to support a growing movement towards new forms of undergraduate research and teaching. McManus and Rubino in 2003 also provided a brief overview of the state of Internet resources available in classics and focused in particular on its impact on pedagogy (McManus and Rubino 2003)

Instead of focusing on the uniqueness of classical computing, Greg Crane proposed instead in 2004 that classical studies no longer needed its own separate computing history. He noted that the use of computing in classical studies was long-standing largely due to the fact that the study of antiquity has always been a "data-intensive" enterprise and that all the reference works and critical editions created by classicists were well suited to an electronic environment. Crane offered his own summary of why classics had for a long time needed its own separate computing history, but his essential argument was that such histories would no longer need to be written:

> There should not be a history of classics and the computer, for the needs of classicists are simply not so distinctive as to warrant a separate " informatics." Disciplinary specialists learning the strengths and weaknesses have, in the author's experience, a strong tendency to exaggerate the extent to which their problems are unique and to call for a specialized, domain-specific infrastructure and approach (Crane 2004).

The needs of classicists in 2004, Crane argued, were not so different from those of other disciplines, and classicists would need to learn to adapt the tools of other disciplines. While Melissa Terras has also recently agreed with Crane's call for classicists to work in an interdisciplinary manner and adapt the computational advances and infrastructure from other disciplines, she also made the cogent point that there are logistical and personal issues of disciplinarity that need to be considered when pursuing cross-disciplinary work (Terras 2010), a issue that has received little research.[9] Disciplinarity has presented digital classicists with two key challenges Terras argued, the first of which is the difficulty of "forging an identity and gaining recognition" for their work within the traditional discipline of classics, and secondly, for those scholars who go beyond traditional classics, there will be the difficulties of engaging with experts in various computer science disciplines:

> Classicists using digital technologies in their research are regularly at the forefront of research in digital humanities, given the range of primary and secondary sources consulted, and the array of tools and techniques necessary to interrogate them. However, to adopt new and developing techniques, and to adopt and adapt emergent technologies, the Digital Classicist has to work in the interdisciplinary space between Classics and computing science (Terras 2010, pg. 178).

Two projects that Terras believed illustrated the interdisciplinary vision necessary to pursue successful work in digital classics were eSAD and VERA, both of which have required large-scale interdisciplinary teams to conduct their work.

[9] Terras cites (Siemens 2009) as one of the few research articles in this area.

Balancing the needs of individual disciplines such as classics within a larger scholarly cyberinfrastructure that is still useful across disciplines is a challenge echoed by many humanities cyberinfrastructure projects as shall be discussed later in this report. Before we move to questions of infrastructure, whether for classics or for the humanities as a whole, however, the next section will look at some large multidisciplinary classical digital libraries that exist, the services they provide, and the state-of-the-art in potential services they might develop.

Multidisciplinary Classical Digital Libraries: Advanced Technologies and Services

The extensive interest in the language, literature and history of the ancient world is clearly evidenced by the large number of resources available online. Digital collections of classical texts (particularly in Greek and Latin) abound online and have been created both by enthusiasts and by academics. For Latin, some of the larger collections are Corpus Scriptorum Latinorum, IntraText, Lacus Curtius, and the Latin Library, for Latin and Greek there is the Bibliotheca Augustana and the Internet Classics Archive, and for English translations of Greek mythology there is the Theoi E-Text collection.[10] Typically these collections involve digital provision of texts that have been typed in manually and the main form access is provided by a browseable list of authors and works. Basic services such as keyword searching are also typically implemented.

A way perhaps of distinguishing between a digital collection and a digital library, is that a library provides a variety of *services* in addition to an organized collection of objects and texts. Candela et al. (2007) proposed that the definition of digital libraries has expanded recently because "generally accepted conceptions have shifted from a content-centric system that merely supports the organization and provision of access to particular collections of data and information, to a person-centric system that delivers innovative, evolving, and personalized services to users." Focusing on this concept of services, this section shall provide brief overviews of a number of multidisciplinary classical digital libraries that are large in scope and also provide specialized services to their users including the Cuneiform Digital Library Initiative (CDLI),[11] Perseus Digital Library (PDL)[12] and the TLG.[13] It shall then look at a number of technologies that could be used to provide advanced services for classical digital libraries in general.

The CDLI is a joint project of the University of California-Los Angeles and the Max Planck Institute for the History of Science, and according to its website, "represents the efforts of an international group of Assyriologists, museum curators and historians of science to make available through the internet the form and content of cuneiform tablets dating from the beginning of writing, ca. 3350 B.C., until the end of the pre-Christian era." While this group estimates that there are about 500,000 documents available in private and public collections, the CDLI currently provides online access to more than 225,000 that have been cataloged. The CDLI maintains an extensive website with a full list of included collections, educational resources, a list of related publications,[14] project partners, tools and resources, and extensive documentation regarding how data is entered and transliterated and how the online catalog was created. Access to the collection is supported by both a general and advanced search option. The basic search interface supports searching transliterations, the catalog or both simultaneously, while the advanced search also supports searching by publication information, physical information, text content (with language limits), provenience and chronology. The record for each document includes extensive catalog information, hand-drawn images or digital images and extensive transliterations.[15]

The PDL, currently in version 4.0, first began in 1985, and it has evolved from CD-Roms to the current version of its online digital library. While its flagship collection is a classical collection that includes a large number of

[10] For a list of these URLs plus some other selected collections, see http://delicious.com/AlisonBabeu/digital_collections
[11] http://cdli.ucla.edu/
[12] http://www.perseus.tufts.edu/hopper/
[13] http://www.tlg.uci.edu/
[14] These publications include the CDLJournal, the CDLBulletin and CDLNotes. In general, many of these scholarly and peer-reviewed publications (all of which are freely available online) represent fairly traditional research that has been made possible through the availability of the online collection but there is also research that has made use both of the CDLI and computational techniques, see for one example (Jaworski 2008)
[15] http://cdli.ucla.edu/search/result.pt?id_text=P020005&start=0&result_format=single&-op_id_text=eq&size=100

Greek and Latin texts (currently 8,378,421 Greek words and 8,696,429 Latin words), multiple English translations, and various reference works such as commentaries, histories and lexicons, it also has other collections in Arabic, Old Norse, and English. An Art and Archaeology browser also provides access to an extensive collection of images. All of the texts in the PDL are encoded in TEI-XML and the majority of them as well as the source code or "hopper" that runs the production digital library can be downloaded. [16] For the majority of Latin and Greek collections, the PDL presents an online reading environment that provides parallel texts (Greek and English, Latin and English) where a user can read a Greek or Latin text aligned with a public domain English translation. Each text in the various collections is also extensively hyperlinked to relevant entries in lexicons, dictionaries and commentaries. The classics collection can be browsed by author and a number of sophisticated searching options are available including multi-lingual searching (English, Greek, Latin, Old, English, German and Old Norse) of the whole digital library, individual collections or individual texts. Named entity searching (people, places, and dates) is also available. The PDL also offers several useful linguistic tools including an English lookup of words in Greek, Latin and Arabic based on their English definitions, a vocabulary tool and a word study tool. While a full overview of the history, services and collections available at the PDL is beyond the scope of this review, a full list of publications (with most available for download) is available online. [17]

The TLG is arguably the best-known digital library in classics and was founded in 1972 (Brunner 1991). Based at the University of California-Irvine, the TLG has "collected and digitized most texts written in Greek from Homer (8 c. B.C.) to the fall of Byzantium in AD 1453 and beyond." The main goal of the TLG "is to create a comprehensive digital library of Greek literature from antiquity to the present era." The TLG was first available on magnetic tapes and then on CD-Rom, and since 2001 it has been available online by subscription with its own search engine. The TLG contains more than "105 million words from over 10,000 works associated with 4,000 authors." The corpus can be browsed by author or searched by author, work title or TLG number and recently started supporting lemmatized searching.

Large digital libraries of classical materials such as those listed above are highly curated and offer very specialized services. As mass digitization projects such as Google Books and the Open Content Alliance continue to put overwhelming materials of Greek, Latin, and other historical languages such as Sanskrit online, ways to scale these services to meet the needs of larger collections are increasingly important. In addition, there are many reference tools online such as bibliographies and directories that will likely need to adapt to the scale of million book classical digital libraries. The rest of this section will briefly focus on some of the important types of cross-disciplinary services, tools and technologies currently available for digital classics and how such domain tools might be of use in a classical cyberinfrastructure.

Bibliographies/Catalogs/Directories

The wealth of research and finding tools for digital classics matches the large number of digital collections and digital libraries of classical materials available. This subsection will provide a brief overview of some of these tools.

The digital environment has provided a useful way of updating and providing access to bibliographies. One of the oldest bibliographies online is ABZU, [18] which since 1994 has been providing a guide to "networked open access data relevant to the study and public presentation of the Ancient Near East and the Ancient Mediterranean world." ABZU is managed by Charles E. Jones, the head librarian at ISAW, and resources include websites as well as open access electronic publications. The collection can either by browed by author or searched by a variety of criterion.

Another significant online bibliography is the "Checklist of Greek, Latin, Demotic and Coptic Papyri, Ostraca and Tablets," [19] which has been created to "provide for scholars and librarians a ready bibliography of all

[16] http://www.perseus.tufts.edu/hopper/opensource/download
[17] http://www.perseus.tufts.edu/hopper/about/publications
[18] http://www.etana.org/abzu/
[19] http://scriptorium.lib.duke.edu/papyrus/texts/clist.html

monographic volumes, both current and out-of-print, of Greek, Latin, Demotic and Coptic documentary texts on papyrus, parchment, ostraca or wood tablets." Periodical articles are not included in this bibliography. Although this bibliography was last updated in 2008, this is a useful tool that also provides a convenient source of the abbreviations used for these monograph collections.

A major bibliography regarding the reception of classical texts[20] by later authors is "The Traditio Classicorum"[21] which has been created by Charles H. Lohr of the University of Freiburg. Available in German or English, this website contains "a bibliography of secondary literature concerning the fortuna of classical authors to the year 1650." The bibliography is arranged by the common Latin names of authors (whether the author wrote in Arabic, Greek or Latin) and the entries for authors have then been divided into general works with specific titles arranged chronologically.

The LDAB (Leuven Database of Ancient Books)[22] now a participant in the Trismegistos portal also supports research into the reception of classical texts. This database collects basic information on ancient literary texts or works (rather than documents) from the 4th century B.C. to 800 A.D. and includes over 3600 "anonymous" texts. According to the website:

> Text editions by classical philologists and patristic scholars are usually based upon medieval manuscripts, dating many centuries after the work in question was first written down and transmitted by copies from copies from copies. Here the user will find the oldest preserved copies of each text. At the same time he will get a view of the reception of ancient literature throughout the Hellenistic, Roman and Byzantine period: which author was read when, where and by whom throughout Antiquity.

Due to the focus of the LDAB on books, this project has excluded documentary texts and references to inscriptions. The database has a variety of advanced searching features including publication, editor, catalogues, ancient author, book, century, date, provenance, nome/region, material (papyrus, parchment), bookform, language/script, script type (Greek, Latin, Demotic, Coptic), etc. For example, searching on the author Herodotus provides a list of documents that have discussed or made references to his history. The LDAB provides an excellent way to study the reception of various classical authors and provides excellent links into various papyri collections.

Rather than providing a bibliography on a particular topic, Pinax Online[23] offers an annotated list of links to online bibliographies regarding the Ancient Greek world. This website is maintained by Marc Huys of the Department of Classical Studies, Katholieke Universiteit Leuven and links are included to general bibliographies of the Greek world, bibliographies for individual Greek authors, and thematic bibliographies (literature, linguistics, mythology and religion, history, and archaeology).

In addition to bibliographies, there are a number of important online catalogues for finding digital classical and medieval materials. LATO, or Library of Ancient Texts Online,[24] has the goal of providing the Internet's "most thorough catalogue of online copies of ancient Greek texts, both in Greek and in translation." This website does not host any actual texts but instead maintains a set of extensive links to Greek texts and their translations on other sites such as Bibliotheca Augustana, Perseus, Project Gutenberg and Theoi Greek Mythology. The links to texts are organized in alphabetical order by author with a list of their works then organized by title. This catalog also includes links to many fragmentary authors

Another extremely useful resource is the "Catalogue of Digitized Medieval Manuscripts"[25] that has been created by the University of California-Los Angeles. This catalog attempts to provide a straightforward way of discovering medieval manuscripts on the web and is still labeled as a "work in progress." The catalogue currently contains over 3000 manuscript descriptions that can be searched by keyword, and the catalogue can

---

[20] Another project that explores the continuing reception of classical texts, albeit with a focus on classical drama, is the Archive of Performances of Greek and Roman Drama Database (http://www.apgrd.ox.ac.uk/database.htm), which "offers information on more than 9000 productions of ancient Greek and Roman drama performed internationally on stage, screen, and radio from the Renaissance to the present day."
[21] http://www.theol.uni-freiburg.de/forschung/projekte/tcdt/index_en.html
[22] http://www.trismegistos.org/ldab/index.php
[23] https://perswww.kuleuven.be/~u0013314/pinaxonline.html#Specifiek
[24] http://sites.google.com/site/ancienttexts/
[25] http://manuscripts.cmrs.ucla.edu/languages_list.php

also be browsed by location, shelfmark, author, title or language (including Arabic, Greek, and Latin). Each manuscript description also contains a link to the digitized manuscript.

A number of specialized research portals and directories to classical materials that also encompass catalogues are also available online. The German website KIRKE (Katalog der Internetressourcen für die Klassische Philologie) [26] has been created by Ulrich Schmitzer and includes an extensive online directory of resources on the Internet for classicists. Another useful resource is SISYPHOS, [27] a searchable directory of over 2100 cataloged Internet resources created by UB Heidelberg. This resource provides access to "Classical Archaeological, Ancient Near Eastern and Egyptological websites" including subject portals, databases of images, mailing lists and discussion forums. One useful feature is that this site provides a "full text search" of the websites it has cataloged not just a search of the metadata of the resource descriptions. Interfaces to this website are available in English and German.

One of the most extensive resources available that covers multiple disciplines within classics is Propylaeum: A Virtual Library of Classical Studies. [28] This subject portal [29] encompasses eight areas of Classical studies: Egyptology, Ancient History, Ancient Near Eastern Studies, Byzantine Studies, Classical Archaeology, Classical Philology, Medieval and Neo-Latin Philology and Pre- and Early History. The entire collection of multi-disciplinary resources can be searched at one time or a full list of resources for individual disciplines can be browsed. Each subject has its own sub-portal that includes a definition of the subject, a list of specialist library catalogues, new acquisitions for partner collections, list of both traditional and e-journals, a list of subject databases and digital collections, links to more general Internet resources, and a list of specialized academic and research services. Over six academic and museum project partners are involved in both developing and maintaining this academic portal including the Bavarian State Library (Munich) and the Institute of Classical Philology at the Humboldt University, Berlin.

In addition to online bibliographies, catalogs and resource directories, there are also a number of online bibliographical research databases for classics. L'Année Philologique (APh) is considered to be the preeminent research tool for secondary literary in classics and is published by the Société Internationale de Bibliographie Classique (overseen by Eric Rebillard) along with the APA and the Database of Classical Bibliography (managed by Dee L. Clayman), with support from both the Centre National de la Recherche Scientifique and the NEH. The goal of the APh is to annually collect scholarly works relating to all aspects of Greek and Roman civilization not just within classics but also from its "auxiliary" disciplines of archaeology, epigraphy, numismatics, papyrology and palaeography. The APh has collaborating teams in France, Germany, Italy, Spain, Switzerland and the United States, and item abstracts and searching in the database are available in all of theses languages. Every year a printed volume is created of the entire bibliography and this is then uploaded into the database where it can be searched by author of the scholarly work, full text (of the abstract), ancient author and text that is referenced, and subject and discipline. Currently, the APh includes bibliographic records for works published through 2007 and is available through subscription.

Two freely available bibliographic databases also exist for classical studies. GNOMON ONLINE [30] is maintained by Jürgen Malitz of Catholic University Eichstatt-Ingolstadt and Gregor Weber of University Augsburg. Its main interface is in German and the bibliographic metadata in the database can either be searched or browsed by a thesaurus. TOCS-IN [31] provides access to the tables of contents from around 185 journals (and thus over 45,000 articles) in classics, Near Eastern Studies, and religion, both in a text format and through a Web program. According to the website access is provided to full text articles about 15% of the time. TOCS-IN is an entirely volunteer project that first began to archive table of contents in 1992 and is currently managed

[26] http://www.kirke.hu-berlin.de/ressourc/ressourc.html
[27] http://vifa.ub.uni-heidelberg.de/sisyphos/servlet/de.izsoz.dbclear.query.browse.Query/domain=allg//lang=en?querydef=query-simple
[28] http://www.propylaeum.de/interdisciplinary.html?L=1
[29] It also searches both KIRKE and SISYPHOS.
[30] http://www.gnomon.ku-eichstaett.de/Gnomon/en/Gnomon.html
[31] http://www.chass.utoronto.ca/amphoras/tocs.html

by PMW Matheson in Toronto. Around 80 volunteers from 16 countries contribute tables of contents to this service. TOCS-IN can be either searched or browsed.

Finally, DAPHNE (Data in Archaeology, Prehistory and History on the Net)[32] is a freely available portal that provides a single point of access to subject-oriented bibliographic databases in prehistory, protohistory, archaeology, and the sciences of antiquity, until around 1000 A.D. DAPHNE combines resources from three French databases: BAHR (Bulletin Analytique d'histoire Romaine), FRANCIS, and FRANTIQ-CCI. Users of DAPHNE can search across the bibliographic records in these databases in Dutch, English, French, German, Italian and Spanish.

Document Analysis, Recognition and OCR for Historical Languages

While the major classical digital libraries listed above support searching across their collections in the original languages such as Latin and Greek, implementing such techniques in a scalable manner is far more challenging. Conventional optical character recognition (OCR) systems have a limited ability to work with historical languages such as Latin, but languages such as Greek and Sanskrit are even more problematic. Special document recognition and analysis systems have been developed to deal with many of the issues these languages present and the research literature on the topic of historical document analysis and recognition is extensive.[33] Some research is starting to explore how technologies that have been developed might be used to solve problems across historical languages or types of object whether it be an inscription, papyri, or palimpsest.[34] This section will present some brief highlights on the use of document analysis and OCR technologies to support information access to documents in classical languages such as Ancient Greek, Latin, Sanskrit, Sumerian and Syriac.

One recent major project that has been funded in this area is "New Technology for Digitization of Ancient Objects and Documents," a joint project of the Archaeological Computing Research Group (ACRG) and the School of Electronics and Computer Science (ECS), Southampton and the Centre for the Study of Ancient Documents (CSAD), Oxford, the Cuneiform Digital Library Initiative (CDLI), Los Angeles-Philadelphia-Oxford-Berlin, and the Electronic Text Corpus of Sumerian Literature (ETCSL), Oxford.[35] This project has received a 12 months Arts and Humanities Research Council (AHRC) grant to "develop a "Reflectance Transformation Imaging (RTI) System for Ancient Documentary Artefacts." The team plans to develop two RTI systems that can be used to capture high quality digital images of documentary texts and archaeological materials, and the initial testing will be conducted on stilus tablets from Vindolanda, stone inscriptions, Linear B and cuneiform tablets.

Other relevant research is being conducted by the IMPACT (Improving Access to Text)[36] project. This project is being funded by the European commission and it is currently exploring how to develop advanced OCR methods for historical texts, particular in terms of the use of OCR in mass digitization processes.[37] While their research is not specifically focused on developing techniques for classical languages, Latin was the major language of intellectual discourse in Europe for almost a century, so techniques adapted for either manuscripts or early printed books would be very useful to classical scholarship and beyond.

Ancient Greek

There has been a limited amount of work that has considered using automatic techniques in the recognition of Ancient or classical Greek. While some recent research has focused on the development of OCR for "Old

[32] http://www.daphne.cnrs.fr/daphne/search.html;jsessionid=5636B7A687E5E01429A1FD79CC88168B
[33] A full review of this literature is beyond the scope of this paper, but for some recent overviews in terms of digital libraries see (Sankar et al. 2006) and (Choudhury et al. 2006).
[34] In fact, a conference to be held in the fall, "Digital imaging of ancient textual heritage: technological challenges and solutions"(http://www.eikonopoiia.org/home.html) explores these issues. Also for a multilingual approach to manuscripts, see (Leydier et al. 2009)
[35] http://www.southampton.ac.uk/archaeology/news/news_2010/acrg_dedefi_main.shtml
[36] http://www.impact-project.eu/home/
[37] For a recent overview of some of the IMPACT project's research see (Ploeger et al. 2009).

Greek" historical manuscripts,[38] little work has explored developing techniques for either manuscripts or printed editions of Ancient Greek texts.

Some preliminary work in developing an automatic recognition methodology for Ancient Greek was detailed in (Stewart et al. 2007). In an initial survey of Greek editions, they found that on average almost 14% of the Greek words on a text page were found in the notes or apparatus criticus. The authors first used a multi-tiered approach to OCR that applied two major post-processing techniques to the output of two commercial OCR packages, ABBYY FineReader (8.0) [39] and Anagnostis 4.1. During this first experiment, character accuracy on simple uncorrected text averaged about 98.57%. Other preliminary experiments with OCR-generated text also revealed that the uncorrected OCR could serve as searchable corpora. Even when working with a mid-19th century edition of Aristotle in a non-standard Greek font, searching of the OCR generated text vs. texts that had been manually typed in, typically provided superior recall because the OCR text included variant readings found in the notes. In a second experiment, the automatic correction of single texts was performed using a list of one million Greek words and the Morpheus Greek morphological analyzer that was developed by the Perseus Digital Library.

A third experiment used the OCR output of multiple editions of the same work to correct one another in a process that involved three steps. First, different editions of a text were aligned by finding unique strings in each, second, if an error word was found in one text a fuzzy search was performed in the aligned parallel text to locate a potential correct form, and third, once error words in a base text were matched against potential ground truth counterparts in the parallel texts, rules generated by the decision tree program (C4.5) were used to determine the more likely variant. The authors found that the parallel text correction rate was consistently higher than the single text correct rate by between 5% and 16%. Baseline character accuracy in this final experiment rose to a level of 99.49%.

The ability to search text variants and to automatically collate various editions of the same work in a digital library through the use of OCR and a number of automated techniques offers up a number of new research opportunities. In addition, as detailed by Stewart et al, their work also provides useful lessons in how curated digital corpora, automated methods and million book libraries can be used to create new more sophisticated digital libraries:

> By situating corpus production within a digital library (i.e., a collection of authenticated digital objects with basic cataloging data), exploiting the strengths of large collections (e.g., multiple editions), and judicious use of practical automated methods, we can start to build new corpora on top of our digital libraries that are not only larger but, in many ways, more useful than their manually constructed predecessors (Stewart et al. 2007).

Further research reported by Boschetti et al. (2009) was informed by the preliminary techniques reported in (Stewart et al. 2007), but also expanded it since this initial work did not include the recognition of Greek accents.

Boschetti et al. (2009) conducted a series of experiments in attempting to create a scalable workflow for outputting highly accurate OCR of Greek text. This workflow utilized progressive multiple alignment of the OCR output of two commercial (Anagnostis, Abbyy FineReader) and one open source OCR engine (OCRopus)[40], a product that was not available when (Stewart et al. 2007) conducted their research. Multiple editions of Athenaeus *Deipnosophistae*, one edition of Aeschylus, and a 1475 edition of Augustine's *De Civitate Dei* were utilized for the OCR experiments. This research determined that the accuracy of single engines was very dependent on the training set created for it, but it also discovered that in several cases OCRopus obtained better results than either commercial option. The highest accuracy level of 99.01%, which was for sample pages from the fairly recent Loeb edition of Athenaeus, was obtained through the use of multiple progressive alignment and a spell-checking algorithm. Accuracy levels on earlier editions of Athenaeus ranged from 94 to 98%. The addition of accents did produce lower character accuracy results that those reported by (Stewart et al.

[38] For an example, see (Ntzios et al. 2007)
[39] http://www.abbyy.com/

2007), but at the same time, accents are an important part of Ancient Greek, and any OCR system ultimately developed for this language will likely need to consider them. This research also demonstrated that OCRopus, an only recently available open source OCR engine, could produce results comparable to expensive commercial products.

While both Stewart et al. (2007) and Boschetti et al. (2009) focused on using OCR to recognize printed editions of Ancient Greek, a variety of both classical scholarship and document recognition research[41] have been conducted on the Archimedes Palimpsest,[42] a 13th century prayer book that contains erased texts that were written several centuries before, including previously "lost" treatises by Archimedes and Hypereides. This manuscript has since been digitized, and the images created of the manuscript pages and the transcriptions of the text are available for download online.[43] Scholars are working with digital images rather than the manuscript itself and a very diverse set of disciplines including palaeography, the history of mathematics and science, and Byzantine liturgists have done extensive work with this palimpsest. Much of the image processing work with the palimpsest has focused on developing algorithms to extract the text of Archimedes in particular from page images. Salerno et al. (2007) used principal component analysis (PCA) and independent component analysis (ICA) techniques to extract "clean maps of the primary Archimedes text, the overwritten text, and the mold pattern present in the pages" from 14 hyperspectral images of the Archimedes. Their main goals were both to provide better access to the text of Archimedes and to develop techniques that can be used in other palimpsest digitization projects. The authors also report that:

> A further aspect of the problem is to partly automate the reading and transcription tasks. This cannot be intended as a substitution of the human experts in a task where they perform better than any presently conceivable numerical strategy, but as an acceleration of the human work (Salerno et al. 2007).

The importance of not replacing expert scholars with systems but rather in developing tools that assist them in their traditional tasks was a theme seen throughout the literature.

Other significant work in the area of providing access to fragile manuscripts has been conduced by the EDUCE (Enhanced Digital Unwrapping for Conservation and Education) Project.[44] This NSF funded project has been working to develop systems that support the "virtual unwrapping and visualization of ancient texts." According to their website,

> The overall purpose is to capture in digital form fragile 3D texts, such as ancient papyrus and scrolls of other materials using a custom built, portable, multi-power CT scanning device and then to virtually "unroll" the scroll using image algorithms, rendering a digital facsimile that exposes and makes legible inscriptions and other markings on the artifact, all in a non-invasive process.

Some of the EDUCE project's image processing techniques have been utilized by the Homer Multitext[45] project as described Baumann and Seales (2009), who presented an application of image registration techniques, or the "process of mapping a sensed image into the coordinate system of a reference image," to the Venetus A manuscript of the *Iliad* used in this project. The Homer Multitext project included 3d scanning as part of its digitization strategy but as the 3D scanning system acquired untextured 3D models a "procedure to register the 2D photography to the 3D scans was performed periodically." It was discovered during one photography session that technical issues had produced a number of images of poor quality. While these images were reshot time constraints prevented re-performing the 3D geometry capture for these pages again. The end result was a number of folios that had two sets of data, a "dirty" image that had registered 3D geometry and a "clean" image with no associated geometry that the project wished to apply digital flattening algorithms too. The main computational problem was thus to determine a means of obtaining a "high-quality deformation of the "clean image" such that the text was in the same position as the "dirty" image" that would then allow them to "apply digital flattening using the acquired corresponding 3D geometry."

---

[41] A palimpsest is a manuscript "on which more than one text has been written with the earlier writing incompletely erased and still visible" (http://wordnetweb.princeton.edu/perl/webwn?s=palimpsest). For a full list of research publications using the Archimedes Palimpsest, see http://www.archimedespalimpsest.org/bibliography1.html

[42] http://www.archimedespalimpsest.org/

[43] http://archimedespalimpsest.net/

[44] http://www.stoa.org/educe/

[45] http://chs.harvard.edu/wa/pageR?tn=ArticleWrapper&bdc=12&mn=1169

The image registration algorithm that Baumann and Seales developed was successful and while the technical details are beyond the scope of this paper, the authors rightly conclude that:

> High-resolution, multispectral digital imaging of important documents is emerging as a standard practice for enabling scholarly analysis of difficult or damaged texts. As imaging techniques improve, documents are revisited and re-imaged, and registration of these images into the same frame of reference for direct comparison can be a powerful tool (Baumann and Seales 2009).

The work of the EDUCE project illustrates how the state-of-the-art is currently being used to provide new levels of access to valuable and damaged manuscripts.

Latin

Between the extensive digitization of cultural heritage materials such as manuscripts and the large number of Latin texts that are becoming available through massive digitization projects, techniques for improving access to these materials is an area of growing research that will be briefly examined in this subsection.

A variety of approaches have been explored for improving access to Latin manuscripts. Leydier et al. (2007) explored the use of word-spotting to improve information retrieval of textual data in primarily Latin medieval manuscript images:

> In practice, word-spotting consists in retrieving all the occurrences of an image of a word. This template word is selected by the user by outlining one occurrence on the document. It results in the system proposing a sorted list of hits that the user can prune manually…. Word-spotting is based on a similarity or a distance between two images, the reference image defined by the user and the target images representing the rest of the page or all the pages of a multi-page document. Contrary to text query on a document processed by OCR, a word-image query can be sensitive to the style of the writing or the typography used. This technique is used when word recognition cannot be done, for example on very deteriorated printed documents or on manuscripts (Leydier et al. 2007).

The main drawback to this approach as reported by the authors is that a user has to select a keyword in a manuscript image (typically based on an ascii transcript) as a basis for further image retrieval, limiting their approach to retrieval of other images by word only.

Another approach presented by Edwards et al. (2004) trained a generalized Hidden Markov Model (gHMM) on the transcription of a Latin manuscript to get both a transmission model and one example each for 22 letters to create an emission model. Their transition model for unigrams, bigrams, and trigrams was fitted using the Latin Library's electronic version of Caesar's the *Gallic Wars* and their emission model was trained on 22 glyphs taken from a 12th century manuscript of Terence's *Comoediae*. In contrast to the approach of Leydier et al., the authors argued that word spotting was not entirely appropriate for a highly-inflected language such as Latin:

> Manmatha et al…introduce the technique of "word spotting," which segments text into word images, rectifies the word images, and then uses an aligned training set to learn correspondences between rectified word images and strings. The method is not suitable for a heavily inflected language, because words take so many forms. In an inflected language, the natural unit to match to is a subset of a word, rather than a whole word, implying that one should segment the text into blocks — which may be smaller than words — while recognizing (Edwards et al. 2004).

In their model, they chose not to transition word to word transition probabilities since word order in Latin is highly arbitrary. The method developed had reasonable accuracy, 75% of letters were correctly transcribed and relatively strong searching ability was reported.

Some research with document analysis of Latin manuscripts has focused on assisting palaeographers. The discipline of palaeography will be explored further in its subsection, but in general, palaeography studies the writing style of ancient documents.[46] Moalla et al. (2006) conducted automatic analysis of the writing styles of ancient Latin manuscripts from the 8th to 16th centuries and focused on the extraction of "sufficiently discriminative features" in order to be able to differentiate between a sufficiently large number of Latin writings. A number of problems complicated their image analysis including the complexity of the shapes of letters, the existence of hybrid writing styles, bad manuscript quality, overlapping of lines and words, and poor quality

[46] An excellent resource for exploring ancient writing systems is Mnamon: Ancient Writing Systems in the Mediterranean (http://lila.sns.it/mnamon/index.php?page=Home&lang=en) that not only provides extensive descriptions on various writing systems but also includes selected electronic resources.

manuscript images. Their discriminant analysis of fifteen Latin classes only achieved a classification accuracy rate of 59% in their first iteration but the elimination of four classes that were not statistically well-represented increased the rate to 81%.

Another key area of technology research is in the development of techniques for digitizing and searching incunabula, or early printed books, a large number of which were printed in Latin. One major project in this area is CAMENA—Latin Texts of Early Modern Europe,[47] a project hosted by the university of Mannheim. Their digital library currently includes five collections: a collection of Neo-Latin poetry composed by German authors available as images and machine readable texts, a collection of Latin historical and political writing from early modern Germany, a reference collection of dictionaries and handbooks from 1500-1750 that helps provide a reading environment, a corpus of Latin letters written by German scholars between 1530 and 1770, and a collection of early printed editions of Italian Renaissance humanists born before 1500. This project also includes the Termini and Lemmata databases, two projects that are now part of the larger eAQUA project. The wealth of Neo-Latin materials online is well documented by the "Philological Museum: An Analytic Bibliography of On-Line Neo Latin Texts,"[48] an extensive website created by Dana F. Sutton of the University of California-Irvine that since 1999 has served as an "analytic bibliography of Latin texts written during the Renaisance and later that are freely available to the general public on the Web" and it includes over 33,960 records.

Digitizing incunabula or early modern printed books, however, is not an easy task, and includes a number of challenges outlined by Schibel and Rydberg-Cox (2006) and Rydberg-Cox (2009).

> Incunabula, or books printed before 1500, are extremely difficult and expensive to convert to digital form. The primary challenges arise from the use of non-standard typographical glyphs based on medieval handwriting to abbreviate words. Further difficulties are also posed by the practice of inconsistently marking word breaks at the end of lines and reducing or even eliminating spacing between some words (Rydberg-Cox 2009).

In addition, such digitized texts are often presented to a modern audience only after an extensive amount of editing and annotation has occurred, a level of editing that is not scalable to million book libraries.

Schibel and Rydberg-Cox argued that good bibliographic description is required for this historical source material (ideally so that such collections can be sorted by period, place, language, literary genre, publisher and audience), particularly since many digitized texts will often be reused in other contexts. A second recommendation made by Schibel and Rydberg-Cox (2006) is the need to identify at least basic structural metadata for such books (front, body, back, etc.) or to create a rough table of contents that provides a framework by which to make page images available. They suggested that such structural metadata would support new research into traditional questions of textual influence for researchers who could use automatic text similarity measures to recognize text families and trace either the influence of major authors or the purposes of a given document. Despite such new opportunities, a number of problems remain, for an initial analysis by the authors of digital libraries of page images of early modern books revealed that page images produced were often inaccurate or inadequate, OCR tools were not yet flexible enough to produce transcriptions, and automated tagging and linking is far more difficult with "pre-standardized language."

Schibel and Rydberg-Cox also concluded, however, that the greatest challenge faced in providing access to early modern books, was that linguistic tools for Early Modern Latin are considerably underdeveloped:

> Aside from the issues outlined above, two major challenges face humans and computers alike. First, we have no comprehensive dictionary of Neo-Latin. Readers must cope with neologisms or, often much harder to decipher, idioms and turns of expression of particular groups. Second, aside from morphological analyzers such as Morpheus – the Latin morphological analyzer found in the Perseus Digital Library – we have few computational tools for Latin. Even Morpheus does not use contextual clues to prioritize analyses, and we are not aware of any substantive work on named entity recognition in Latin. We do not yet have mature electronic authority lists for the Greco-Roman world, much less the people, places, etc. of the early modern period (Schibel and Rydberg Cox 2006).

[47] http://www.uni-mannheim.de/mateo/camenahtdocs/camena.html
[48] http://www.philological.bham.ac.uk/bibliography/

Some of the issues listed here, such as the development of linguistic tools for early Modern Latin have received some further research in the last four years, including (Reddy and Crane 2006). In their work, Reddy and Crane (2006) tested the abilities of the commercial OCR ABBY FineReader and the open source document recognition system GAMERA [49] to recognize glyphs in early modern Latin documents. They found that after extensive training GAMERA could recognize about 80% of glyphs while FineReader could recognize about 84%. In order to improve the character recognition output, they recommended the use of language modeling for future work.

Rydberg Cox (2009) also explored some of the computational challenges in creating a corpus of early Modern Latin and reported on work from the NEH Project, "Approaching the Problems of Digitizing Latin Incunables." The primary aim of this project was to examine the "challenges associated with representing in digital form the complex and non-standard typefaces used in these texts to abbreviate words" a practice that was done in imitation of medieval handwriting practice. Such features of early typography occurred at varying rates in different books Rydberg-Cox noted, but they do appear so frequently that no digitization project can fail to consider them. This issue was also faced by the Archimedes Digital Library [50] project that when digitizing texts published between 1495 and 1691 discovered between three and five abbreviations on every printed page. An important point thus made by Rydberg Cox was that when digitizing early modern books, a project needs to determine how much functionality users will require from a digital facsimile and how much human intervention will be required to create it.

In analyzing these questions, Rydberg-Cox defined five basic possible approaches: 1) Image books with simple page images; 2) "Image books with minimal structural data"; 3) "Image front transcriptions" such as those found in the Making of America [51] project, or where the user is presented with page images that have uncorrected OCR that can be searched but is never seen by the user; 4) Transcriptions (generally marked up in XML) that have been carefully edited and tagged; and finally, 5) Scholarly and critical editions. Ultimately the project decided to create sample texts in all of these genres except that of the scholarly critical edition due to the cost of creating such editions. This decision to digitize the text rather than just provide page images with limited OCR, raised its own series of issues including the need to manually photograph rather than scan pages and how to address characters and glyphs that could not be represented by Unicode. They had to create a method that could be used by data entry contractors to represent characters as they typed up texts, and the first step was to create a catalog of all the brevigraphs that appeared in the printed books, which assigned a unique entity identifier to each non-standard character that data entry personnel could then use to represent the glyph.

In addition to creating this catalog, a number of computational tools were created to assist the data entry operators:

> Because the expansion of these abbreviations is an extremely time-consuming and painstaking task, we developed three tools to facilitate the tagging process. These tools suggest possible expansions for Latin abbreviations and brevigraphs, help identify words that are divided across lines, and separate words that are joined as the results of irregular spacing. All three programs can return results in HTML for human readability or by XML in response to remote procedure call as part of a program to automatically expand abbreviations in these texts (Rydberg Cox 2009).

Another important point raised by Rydberg-Cox was that while the project needed to develop tools such as this, if such tools were shared in a larger infrastructure, they could then be *reused* by the numerous projects out there digitizing Latin books. Ultimately, Rydberg-Cox concluded that this work illustrated that a large-scale project that created image-front editions (e.g. that used uncorrected data that was manually typed to support searching rather than uncorrected OCR) could be affordably managed. In their own workflow, they found the most significant expense was in having human editors tag abbreviations and a second editor proofread work.

Nonetheless, Rydberg-Cox convincingly argued that a certain level of transcription is typically worth the cost for it provides better searchability, and even more importantly, supports automatic hypertext and linking to dictionaries and other reading support tools. Such tools can help students and scholars read texts in Greek and

[49] http://gamera.informatik.hsnr.de/
[50] http://archimedes.fas.harvard.edu/
[51] http://moa.umdl.umich.edu/

Latin without expert knowledge of such languages, and they are particularly important for early modern books since many of these books have never been translated. Furthermore, Rydberg-Cox noted that larger collections of lightly edited text rather than small collections of closely edited texts or critical editions often reach far larger audiences. In addition, this model does not preclude the development of critical editions, for as long as the images and transcriptions are made available as open content they can be reused by scholars in support of making their own editions.

In contrast to utilizing digitized images and typed in transcriptions, recent research reported by Simone Marinai (Marinai 2009) explored the use of automatic text indexing and retrieval methods to support information retrieval from early modern books. She tested her methods on the Latin Gutenberg Bible and reported the same problems as Schibel and Rydberg-Cox, namely the high density of text on each page, the limited spacing among words, and most importantly, the use of many abbreviations and ligatures. She noted that such issues limit not just automatic techniques but human reading as well. The Gutenberg Bible alone included 75 types of ligatures, with 2 dense columns of text per page, with each containing 42 lines. The methodology proposed, Marinai hoped, would support information retrieval beyond this one text:

…our aim is not to deal only with the Gutenberg Bible, but to design tools that can process early printed books, that can adopt different ligatures and abbreviations. We therefore designed a text retrieval tool that deals with the text in a printed document in a different way, trying to identify occurrences of query words rather than recognizing the whole text (Marinai 2009).

Instead of word segmentation, Marinai's technique extracted "Character Objects" from documents that were then clustered together using Self Organizing Maps so that "symbolic" classes could be assigned to indexed objects. User query terms were selected from "one word" images in the collection that were then compared against "indexed character objects with a Dynamic Time Warping (DTW) based approach." This "query by example" approach did face one major challenge in that it could not find occurrences of query words that were printed with different ligatures.

As this subsection indicates, the development of tools for the automatic recognition and processing of Latin is a research area that still has many open challenges and questions.

Sanskrit

The issues involved in the digitization of Sanskrit texts and the development of tools to both study and present them online are so complicated that an annual international Sanskrit computational linguistics symposium was established in 2007.[52] This subsection will provide an overview of some of the major digital Sanskrit projects and current issues in digitization.

The major digital Sanskrit project online is The Sanskrit Library a "digital library dedicated to facilitating education and research in Sanskrit by providing access to digitized primary texts in Sanskrit and computerized research and study tools to analyze and maximize the utility of digitized Sanskrit text."[53] The Sanskrit Library is part of the International Digital Sanskrit Library Integration project that seeks to connect various Sanskrit digital archives and tool projects as well as to establish encoding standards, enhance manuscript access, and develop OCR technology and display software for Devanagari text. On an individual basis, the Sanskrit library supports philological research and education in Vedic and Classical Sanskrit language and literature and provides access to Sanskrit texts in digital form. The Sanskrit Library currently contains independent study Sanskrit readers, grammatical literature, morphological software, instructional materials, and a digital version of W. D. Whitney's *The Roots, Verb-Forms, and Primary Derivatives of the Sanskrit Language*. Their current areas of research include "linguistic issues in encoding, computational phonology and morphology, OCR for Indic scripts, and markup of digitized Sanskrit lexica." Free access to this library is provided but users must register.

[52] http://www.springerlink.com/content/p665684g40h7/?p=967bbca4213c4cb6988c40c0e3ae3a95&pi=0
[53] http://sanskritlibrary.org/

Another major scholarly online collection of Sanskrit is the Digital Corpus of Sanskrit (DCS),[54] which provides access to a searchable collection of lemmatized Sanskrit texts and to a partial version of the database of the SanskritTagger software. SanskritTagger is a "part-of-speech (POS) and lexical tagger for post-Vedic Sanskrit" and it is able to analyze unprocessed digital Sanskrit text both lexically and morphologically.[55] The DCS was automatically created from the most recent version of the SanskritTagger database and the corpus was chosen by the software creator Oliver Hellwig (the website notes that this corpus had made no attempt to be exhaustive). The DCS has been designed to support research in Sanskrit philology and it is possible to search for lexical units and their collocations from a corpus of 2,700,000 words.

A variety of research has been conducted into the development of tools for Sanskrit and this subsection will only briefly review some of this work. The need for digitized Sanskrit lexicons[56] as part of a larger computational linguistics platform is an area of research for the Sanskrit Library, and this issue has received substantial attention in (Huet 2004). This article provided an overview of work to develop both a Sanskrit lexical database and various automatic tagging tools in order to support a philologist:

> The first level of interpretation of a Sanskrit text is its word-to-word segmentation, and our tagger will be able to assist a philology specialist to achieve complete morphological mark-up systematically. This will allow the development of concordance analysis tools recognizing morphological variants, a task which up to now has to be performed manually (Huet 2004).

Huet also asserted that the classical Sanskrit corpus is quite extensive and presents computational linguistics with many analytical challenges.

In addition to the challenges Sanskrit presents for developing computational tools, the features of the language itself make the creation of critical editions very difficult. As Csernel and Patte (2009) explain, a "critical edition" must take "into account all the different known versions of the same text in order to show the differences between any two distinct versions."[57] The creation of critical editions is challenging in any language, particularly if there are many manuscript witnesses, but Sanskrit presents some unique problems. In this paper, Csernel and Patte present an approach based on paragraphs and sentences extracted from a collection of manuscripts known as the "Banaras" gloss. This gloss was written in the 7th century A.D. and is the most famous commentary on the "notorious" Panini grammar, which was known as the first "generative" grammar and was written around the fifth century B.C. One major characteristic of Sanskrit described by Csernel and Patte is that is "not linked to a specific script" and while the Brahmi script was used for a long time, Devanagari is now the most common. The authors reported that they used the transliteration scheme of Sanskrit for Tex that was developed by Frans Velthius[58] wherein each Sanskrit letter is written using between one and three Latin characters.

One interesting insight provided by these authors was how one problematic feature of Sanskrit texts, namely text written without spaces, was also found in other ancient texts:

> In ancient manuscripts, Sanskrit is written without spaces, and from our point of view, this is an important graphical specificity, because it increases greatly the complexity of text comparison algorithms. One may remark that Sanskrit is not the only language where spaces are missing in the text: Roman epigraphy and European Middle Age manuscripts are also good examples of that (Csernel and Patte 2009).

The solution that the authors ultimately proposed for creating a critical edition of a Sanskrit text involved the lemmatization by hand of one of the two texts, specifically the text of the edition. Alignments between this lemmatized text and other texts then made use of the longest common subsequence (LCS) algorithm. Currently they are still experimenting with their methodology, but the authors also pointed out that the absence of a Sanskrit lexicon limited their approach.

---

[54] http://kjc-fs-cluster.kjc.uni-heidelberg.de/dcs/

[55] For more details on this tagger, see (Hellwig 2007), and for one of its research uses in philology see (Hellwig 2010).

[56] The NEH has recently funded a first step in this direction. A project entitled "Sanskrit Lexical Sources: Digital Synthesis and Revision" will support an "international partnership between the Sanskrit Library (Maharishi University of Management) and the Cologne Digital Sanskrit Lexicon (CDSL) project (Institute of Indology and Tamil Studies, Cologne University) to establish a digital Sanskrit lexical reference work." http://www.neh.gov/news/archive/201007200.html

[57] Further discussion of this issue can be found in the subsection on Digital Editions.

[58] http://www.ctan.org/tex-archive/language/devanagari/velthuis/

As has been cited above, the development of OCR tools that will process Sanskrit scripts is a highly sought after goal. Very little work has been done in this area, but Thomas Breuel recently reported on the use of OCRopus not only to recognize the Devanagari script but also on its application to both primary texts in classical languages and to secondary classical scholarship. As was discussed previously in Boschetti et al. (2009), preliminary work with OCRopus produced promising results with Ancient Greek.

OCRopus is described by Breuel et al. (2009) as an open source OCR system that is designed to be omni-lingual and omni-script and it also advances the state-of-the-art in that new text recognition and layout analysis modules can be easily plugged in and it uses an adaptive and user extensible character recognition module. Breuel acknowledged that there are many challenges to recognizing Devanagari script including the large number of ligatures, complicated diacritics, and the "large and unusual vocabulary used in academic and historical texts" (Breuel 2009). In addition to Sanskrit texts, Breuel made the important point that historical scholarship about Sanskrit and other classical languages is frequently multi-lingual and multi-script and can mix Devanagari and Latin as well as Greek. Breuel thus proposed that OCRopus has a number of potential applications in the field of classical scholarship including the recognition of original documents (written records), original primary source texts (printed editions of classical texts), and both modern and historical secondary scholarship including commentaries and textbooks, and reference works such as dictionaries and encyclopedias.

OCRopus is an open source system that Breuel explained uses a "strictly feed-forward system," an important feature that supports the plug-in of other layout-analysis and text-recognition modules. Other features include the use of only a small number of data types to support reuse, "weighted finite state transducers" (WFSTs) to represent the output of text line recognition, and final output in the hOCR format, which "encodes OCR information in completely standards-compliant HTML files." This open source system can be hosted through a web service, run from the command line or shell scripts, and users can customize how it performs by scripting "the OCR engine in Lua."

The basic stages in using OCRopus involve image preprocessing, layout analysis, text-line recognition and statistical language modeling. Each of these stages offers a variety of customization options that make it particularly useful for historical languages. In terms of text-line recognition in historical texts, the fact that OCRopus has both built-in text line recognizers and the ability to add external text-line recognizers for different scripts is very important, because as Breuel articulated:

> Some historical texts may use different writing systems, since Devanagari is not the only script in historical use for Sanskrit. Scholarly writing on Sanskrit almost always uses Latin script, and Latin script is also used for writing Sanskrit itself, including extended passages. Sanskrit written in Devanagari and Latin scripts also makes use of numerous diacritics that need to be recognized. In addition, IPA may be used for pronunciation, Greek letters may be used for classical Greek quotations, and Greek letters and other special characters may be used for indicating footnotes or other references (Breuel 2009).

The challenge of multi-lingual document recognition is a significant one for classical scholarship that was reported by many digital classics projects. OCRopus has built-in line recognizers for Latin scripts, and unlike many other OCR systems, these recognizers make few assumptions about characters sets and fonts and are instead "trained" on text line input that is then aligned against ground truth data and can the be used to automatically train "individual character shape models." For Devanagari, OCRopus handled diacritics by treating "character+diacritic" combinations as novel characters.

The final processing stage of OCRopus involves language modeling, which in the case of OCRopus is based on WFSTs. These WFSTs allow language models and character recognition alternatives to be "manipulated algebraically" and such language models can be learned from training data or constructed manually. One important use of such models for mixed-language classical texts is that they can be used to automatically identify languages within a digital text. "We can take existing language models for English and Sanskrit and combine them," Breuel explicated, "As part of the combination, we can train or specify the probable locations and frequencies of transitions between the two language models, corresponding to, for example, isolated foreign words within one language, or long quotations" (Breuel 2009).

As this subsection has indicated, the computational challenges of processing Sanskrit is a field that is being actively researched, and indeed, some of the technical solutions may very well be adaptable to other historical languages as well.

Syriac

Document recognition of the Syriac dialect, which belongs to the Aramaic branch of the Semitic languages and flourished between the 3rd and 7th century A.D. (although it continued to be used as a written language through the 19th century), has a relatively small body of research. Nonetheless some texts written in this language can be found in many papyri and manuscript collections.[59] Bilane et al. (2008) have investigated the use of word spotting for handwriting analysis in digitized Syriac manuscripts. They noted that Syriac presented a particularly interesting case because it combines the word structure and calligraphy of Arabic handwriting while also being intentionally written at a tilted angle. Earlier work by Clocksin (2003) has also described methods for the automatic recognition of Syriac handwriting in a collection of manuscripts.

Cuneiform Texts and Sumerian

Cuneiform script is generally considered to be the earliest writing system known in the world and it was used in the Ancient Near East from about 3200 B.C. to about 100 A.D. While the largest number of cuneiform texts represent the Sumerian language, the cuneiform script was adapted for other languages including Akkadian, Elamite and Hittite. Sumero-Akkadian cuneiform is the most common by far and is a complex "syllabic and ideographic writing system, with different signs for the various syllables" (Cohen et al. 2004). There are approximately 1000 different cuneiform signs that form a complex script system where most signs are also "polyvalent" or where they have multiple phonemic and semantic realizations. In addition, additional glyphs have also shown great "palaeographic development" over their three millennia of use (Cohen et al. 2004). Sumerian has also been described as a "language isolate" where no other related languages have been identified and so it lacks resources such as a "standardized sign list and comprehensive dictionary" (Ebeling 2007). These various factors make the digitizing, transliterating and presentation of cuneiform online a complicated task.

As indicated by the size of the previously described CDLI, there are also hundreds of thousands of cuneiform tablets and other texts around the world in both private and public collections. In addition to the CDLI, there are a number of significant digital collections and corpora of cuneiform texts online, and this section will describe them briefly along with relevant literature.

One major project to recently emerge from the CDLI is the Open Richly Annotated Cuneiform Corpus (Oracc).[60] This project has grown out of the CDLI and has utilized technology developed by the Pennsylvania Sumerian Dictionary (PSD).[61] According to its website, ORACC was created by Steve Tinney, Eleanor Robson and Niek Veldhuis and "comprises a workspace and toolkit for the development of a complete corpus of cuneiform whose rich annotation and open licensing support the next generation of scholarly research." In addition to CDLI and PSD, a number of other digital cuneiform projects are also involved in Oracc[62] including Assyrian Empire Builders (AEB),[63] the Digital Corpus of Cuneiform Mathematical Texts (DCCMT)[64] and the Geography of Knowledge in Assyria and Babylonia (GKAB).[65] Oracc has been designed as a "corpus building cooperative" that will provide both infrastructure and technical support for "the creation of free online editions of cuneiform texts." Since Oracc wishes to promote both open and reusable data they recommend that all

[59] http://vmr.bham.ac.uk/Collections/Mingana/part/Syriac/
[60] http://oracc.museum.upenn.edu/index.html
[61] The Pennsylvania Sumerian Dictionary project (http://psd.museum.upenn.edu/epsd/index.html) is based at the Babylonian Section of the University of Pennsylvania Museum of Anthropology and Archaeology. In addition to their work with ORACC and the CDLI, they have also collaborated with the Electronic Text Corpus of Sumerian Literature (ETSCL).
[62] For the full list, see http://oracc.museum.upenn.edu/project-list.html.
[63] http://www.ucl.ac.uk/sargon
[64] http://oracc.museum.upenn.edu/dccmt/
[65] http://oracc.museum.upenn.edu/gkab

participating projects make use of Creative Commons (CC)[66] licensing and all default Oracc projects have been placed under a CC "Attribution-Share Alike" license. Oracc has been designed as a complement to the CDLI and allows scholars to "slice" groups of texts from the larger CDLI corpus and then study those intensively within what they have labeled "projects." Among its various features, Oracc supports multilingual translation support, projects can be turned into Word files, PDFs or books using the "ISO OpenDocument" standard, and data can also be exported in the TEI format. Any cuneiform tablet transliterations that are created within Oracc will also be automatically uploaded to the CDLI.

The Oracc project recognizes six major roles[67] and has developed specific documentation for each: user (a scholar using the Oracc corpora), builder (someone working on texts to help build up Oracc, e.g. lemmatizing or data entry), manager (someone actively managing or administering an Oracc project), developer (someone contributing code to the Oracc project), system administrator, and steerer (senior Oracc users). Significant documentation is freely available for all but the last two roles. Oracc is a growing project and researchers are invited to contribute texts to Oracc through either a *donation* or *curation* model: through the donation model text editions and any additional information are simply sent to Oracc and they install, convert and maintain them (Oracc reserves the right for some minor editing but promises proper identification and credit for all data as well as to identify all revisers of data); through the curation model, the Oracc team helps users to set up their Cuneiform texts as a separate project on the Oracc server and the curator is then responsible for lemmatizing and maintaining their texts (this model also gives the user greater control over subsequent edits to materials).[68] Various web services are also provide to assist those that are contributing corpora to Oracc.

Oracc provides an excellent example of a project that supports reuse of its data through the use of CC licenses, commonly adopted technical standards, and extensive documentation as to how the data is created, stored and maintained. By recognizing different roles of users and designing specific documentation for them, Oracc also illustrates the very different skills and needs of its potential users. Finally, through encouraging two different contribution models (both of which encourage sharing and provide attribution), the Oracc project has recognized that there may be many scholars that wish to share their data but either don't wish to maintain it in the long-term or lack the technical skill to do so.

While both the CDLI and Oracc illustrate that there are many currently existing digital cuneiform projects and thousands of digitized cuneiform tablets with both transliterations and translations online, the need to still digitize thousands of cuneiform tablets and to provide long-term access to them is an ongoing challenge. The importance of using 3D scanning as one possible means of preserving cuneiform tablets was discussed by Kumar et al. (2003). The authors observed that cuneiform documents typically exhibit three- dimensional writing on three dimensional-surfaces, so the Digital Hammurabi[69] project described in this article sought to create high resolution 3D models of tablets not only to preserve them but also to provide better access to scholars. Typically cuneiformists have had two main techniques for representing and archiving cuneiform documents, "2D photography and hand-drawn copies, or autographs." In fact, many such autographs can be found in collections such as the CDLI. These autographs, however, have several disadvantages as outlined by Kumar et al. including the fact that they represent one author's interpretation of the signs on a tablet, they cannot be used for collation, and are not very useful for palaeography. The authors thus conclude that:

> It is no wonder then that we are also seeing a number of recent forays into 3D surface scanning of cuneiform tablets, including by our Digital Hammurabi project …. Accurate, detailed, and efficient 3D visualization will enable the virtual "autopsy" of cuneiform tablets and will revolutionize cuneiform studies, not only by making the world's tablet collections broadly available, but also by limiting physical contact with these valuable and unique ancient artifacts, while at the same time providing redundant archival copies of the originals (Kumar et al. 2003).

---

[66] Creative Commons is a "nonprofit corporation dedicated to making it easier for people to share and build upon the work of others, consistent with the rules of copyright" (http://creativecommons.org/about/) and provides free licenses and legal tools that can be used by creators of intellectual works that wish to provide various levels of reuse of their work, including attribution-only, share-alike, non-commercial and no derivative works.

[67] http://oracc.museum.upenn.edu/doc/

[68] For more on the technical details of the curation model see http://oracc.museum.upenn.edu/contributing.html#curation, for their extensive corpus builder documentation http://oracc.museum.upenn.edu/contributing.html#curation) and for the guide to project management http://oracc.museum.upenn.edu/doc/manager/

[69] http://www.jhu.edu/digitalhammurabi/index.html

The Digital Hammurabi project was founded in 1999 and is based at Johns Hopkins University. According to its website, this project has "pioneered basic research on digitizing ancient cuneiform tablets." Their research work has focused on solving three technological problems: 1) the creation of a standard computer encoding for cuneiform text 2) the creation of comprehensive cuneiform collections and 3) solutions for 3d scanning and visualization of the tablets. As of this writing, the project has successfully invented a "3D surface scanner that scans cuneiform tablets at 4 times the resolution of any comparable technology,"[70] developed algorithms designed for "cuneiform tablet reconstruction and 3D visualization" and has successfully overseen a Unicode adoption of "the first international standard for the representation of cuneiform text on computers"(Cohen et al. 2004)

An article by Cohen et al. (2004) has described some of these algorithms, the development of the encoding standard for cuneiform by the "Initiative for Cuneiform Encoding"(ICE),[71] and iClay,[72] "a cross-platform, Internet-deployable, Java applet that allows for the viewing and manipulation of 2D+ images of cuneiform tablets." At the time ICE was formed, there was no standard computer encoding for cuneiform text and Sumerologists had to create Latin transliterations for cuneiform texts. In order to support "automated cuneiform text processing" Cohen et al. stated that a "simple context-free description of the text provided by a native cuneiform computer encoding" was needed. Consequently, ICE developed a cuneiform sign repertoire that merged the three most important sign lists in the world (all unpublished), which was subsequently adopted by Unicode.

Other research into assisting the effective analysis of cuneiform texts has been conducted by the Cuneiform Digital Palaeography (CDP) Project.[73] The CDP is a joint research project between an inter-disciplinary team at the University of Birmingham and the British Museum and it "aims to establish a detailed palaeography for the cuneiform script." The website notes that while palaeography has long been taken for granted in other disciplines it is in its infancy for Assyriology. This project has constructed an online database that includes digital images of individual cuneiform signs taken directly from the original sources and has only used those sources that can be dated to the reign of particular king and are "broadly provenanced." The CDP database can be either browsed or searched and items that are found can be saved to a clipboard and personal notes can be added. Users can access the database as a guest or they can create a registered account.

In addition to research projects on preserving and digitizing cuneiform texts, there are a number of significant cuneiform databases and corpora that are online.[74] The Database of Neo-Sumerian Texts (BDTNS)[75] has been developed by the Centro de Ciencias Humanas y Sociales of the Consejo Superior de Investigaciones Científicas in Madrid. They have created an open database (registration is required to view the unpublished tablets) that manages over 88,000 administrative cuneiform tablets written in the Sumerian language (c. 74,000 published, and 14,000 unpublished). The tablets are from the Neo-Sumerian period (c. 2100-2000 B.C.) and come primarily from five southern cities of Ancient Mesopotamia. A catalogue for both the database and transliterations of the texts provides a variety of searching options, and full records for tablets include extensive descriptive information, original publication details, a drawing of the tablet or digital image, and a link to the CDLI (as there are records for many of the same tablets in both collection).

The Electronic Text Corpus of Sumerian Literature (ETSCL)[76] is a project of the University of Oxford and provides access to a selection of nearly 400 *literary* compositions from the late third and early second century

[70] For more on this scanner, see (Hahn et al. 2006)
[71] http://www.jhu.edu/digitalhammurabi/ice/ice.html
[72] http://www.jhu.edu/digitalhammurabi/iclay/iclayalert.html
[73] http://www.cdp.bham.ac.uk/
[74] In addition to the larger projects and databases discussed in this subsection, there are also many smaller online exhibitions such as "Cuneiform Tablets: From the Reign of Gudea of Lagash to Shalmanassar III" from the Library of Congress, http://international.loc.gov/intldl/cuneihtml/ and the Nineveh Tablet Collection from the British Museum, http://fincke.uni-hd.de/nineveh/
[75] http://bdts.filol.csic.es/
[76] http://etcsl.orinst.ox.ac.uk/#

B.C.[77] This corpus contains Sumerian texts that have been transliterated and also includes English prose translations[78] and bibliographic information for each text. The ETSCL can be browsed or searched and also includes an impressive of list of over 700 signs that provides sign names, an image of the sign, and the ETSCL values for searching the corpus.[79]

Ebeling (2007) has provided an overview of the development of this corpus and the technical challenges therein. Four features make the ETSCL different from other cuneiform projects, first, it is a corpus of literary Sumerian texts rather than administrative Sumerian texts such as those found in the CDLI and the BDTNS, second, it is a corpus of composition, where many "of the individual documents in the corpus are put together from several copies, often damaged or fragmented, of the same text," third, it provides English translations, and fourth, it is the "only corpus of any Ancient Middle Eastern language that has been tagged and lemmatized" (Ebeling 2007). These literary texts also differ from administrative texts in that they spell out the morphology in detail and also provide a source for cultural and religious vocabulary.

The ETSCL like the CDLI and the BDNTS contains transliterations of Sumerian, where the original cuneiform has been converted into and represented by a sequence of Roman characters. As noted above, however, it also contains English translations, and transliterations and translations are linked at the paragraph level. This supports a parallel reading of the original text and translation. In addition, the entire corpus is marked up in TEI (P4) with some extensions in order to accommodate textual variants and linguistic annotations, which Ebeling admitted "sometimes stretched the descriptive apparatus to the limit." One challenge, however, of presenting the text and transliteration side by side in the ETSCL was that the transliteration was often put together from several fragmentary sources. This was solved by using the tagpair <addSpan> and <anchor> from the TEI. The "type" attribute is used to indicate whether it is a primary or secondary variant and a special format was also developed for encoding broken and damaged texts.

One major advance of the ETSCL for corpus studies is that the transliterations were lemmatized with an automatic morphological parser (developed by the PSD project) and the output was then proofread. While this process took a year it also supports lemmatized searching of the ETSCL and when a user clicks on an individual lemma in the ETSCL it can launch a search in the PSD and vice versa. In sum, the ETSCL serves as a "diachronic, annotated, transliterated, bilingual, parallel corpus of literature or as an all-in-one corpus" (Ebeling 2007). The further development of linguistic analysis and corpus search tools for the ETSCL has also been detailed by Tablan et al. (2006):

> The main aim of our work is to create a set of tools for performing automatic morphological analysis of Sumerian. This essentially entails identifying the part of speech for each word in the corpus (technically, this only involves nouns and verbs which are the only categories that are inflected), separating the lemma part from the clitics and assigning a morphological function to each of the clitics (Tablan et al. 2006).

The authors used the open source GATE (General Architecture for Text Engineering)[80] that has been developed at the University of Sheffield, and found that one of the biggest problems in evaluating the success of their methods was that they lacked a morphological gold standard for Sumerian to evaluate their data against. Many of the challenges thus faced by the ETSCL illustrate some of the common issues faced when creating corpora for historical languages, including a lack of lexical resources and gold standard training and evaluation data, the difficulties of automatic processing, and the need to represent physically fragmented sources.

A recently started literary text project is the SEAL (Sources of Early Akkadian Literature)[81] corpus, which is composed of Akkadian (Babylonian and Assyrian) literary texts from the 3rd and 2nd century B.C. that were

---

[77] While the ETSCL focuses on a specific time period, a related project that it appears has just begun is the Diachronic Corpus of Sumerian Literature (DCSL) (http://dcsl.orinst.ox.ac.uk/), a project which seeks to create a "web-based corpus of Sumerian Literature spanning the entire history of Mesopotamian civilization over a range of 2500 years."

[78] English translations of Cuneiform texts are fairly uncommon but another small collection is eTACT, "Electronic Translations of Akkadian Cuneiform Texts" (http://www.etana.org/etact/). This collection (part of ETANA) provides access to translations of 28 Akkadian texts along with full bibliographic information regarding the original cuneiform text.

[79] http://etcsl.orinst.ox.ac.uk/edition2/signlist.php

[80] http://gate.ac.uk/

[81] http://www.seal.uni-leipzig.de/

documented on cuneiform tablets. Funded by the German Israeli foundation for Scientific Research and Development (G.I.F.), this project's goal is to "compile a complete and indexed corpus of Akkadian literary texts from the 3rd and 2nd Millennia BCE" and they hope that this corpus will form the basis for both a history and a glossary of early Akkadian literature. Around 150 texts are available and they are organized by genre classifications (such as epics, hymns and prayers), and edited texts are downloadable as PDFs. While no images of the tablets are given, those texts that are available as downloads include basic catalog information, transliterations, an English translation, commentary and a full bibliography. There are also various indices to the texts including words, deity and personal names, and geographical names (there are separate indices for epics and incantations).

Another growing research project is the Persepolis Fortification Archive Project (PFA Online),[82] a research project of the University of Chicago's Oriental Institute. Archaeologists from the institute working at Persepolis in the 1930s exposed ruins of the palaces of Darius, Xerxes and their successors and found tens of thousands of clay tablets in a fortification wall. These tablets were administrative records produced around 500 B.C. and this archive is being made available for study through PFA Online. This project makes use of OCHRE the Online Cultural Heritage Research Environment and requires the Java Runtime Environment to be used. Since 2002, the PFA project has captured and edited almost 2000 digital images of Elamite tablets, created and edited high resolution digital images of more than 600 Aramaic tablets among a variety of other work[83] and also created a blog to track their progress.[84]

As indicated by this brief overview of research and online projects, the challenge of working with cuneiform tablets and the languages represented on them is an area of active and growing interest.

Computational Linguistics (Treebanks, Automatic Morphological Analysis, Lexicons)

Computational linguistics[85] has been defined as "the branch of linguistics in which the techniques of computer science are applied to the analysis and synthesis of language and speech."[86] Similarly, natural language processing (NLP) has been described as an "area of computer science that develops systems that implement natural language understanding" and is often considered to be a sub-discipline of computational linguistics.[87] The use of both computational linguistics and NLP have grown enormously in the humanities over the last 20 years, and they have an even longer history in classical computing as was described in the introduction to this review.[88] Bamman and Crane (2009) have argued that both computational linguistics and NLP will form necessary components of any cyberinfrastructure for classics:

> In deciding how we want to design a cyberinfrastructure for Classics over the next ten years, there is an important question that lurks between "where are we now?" and "where do we want to be?": where are our colleagues already? Computational linguistics and natural language processing generally perform best in high-resource languages – languages like English, on which computational research has been focusing for over sixty years, and for which expensive resources (such as treebanks, ontologies and large, curated corpora) have long been developed. Many of the tools we would want in the future are founded on technologies that already exist for English and other languages; our task in designing a cyberinfrastructure may simply be to transfer and customize them for Classical Studies (Bamman and Crane 2009).

---

[82] http://ochre.lib.uchicago.edu/PFA_Online/
[83] http://oi.uchicago.edu/research/projects/pfa/
[84] http://persepolistablets.blogspot.com/
[85] Relatively little work has been done utilizing computational linguistics for historical languages such as Latin and Greek, but for some fairly recent experiments in constructing parsers for Latin see (Sayeed and Szpakowicz 2004) and computational grammars for Latin (Casadio and Lambek 2005).
[86] "computational linguistics plural noun" *The Oxford Dictionary of English* (revised edition). Ed. Catherine Soanes and Angus Stevenson. Oxford University Press, 2005. Oxford Reference Online. Oxford University Press. Tufts University. 12 April 2010 <http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t140.e15724>
[87] "natural-language processing" A Dictionary of Computing. Ed John Daintith and Edmund Wright. Oxford University Press, 2008. Oxford Reference Online. Oxford University Press. Tufts University. 12 April 2010 <http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t11.e6410>
[88] For some recent examinations of the potential of both computational linguistics and NLP for the humanities, see (de Jong 2009) and (Lüdeling and Zeldes 2007).

This section will briefly look at three specific applications from the area of computational linguistics and NLP in terms of services for digital classics as a whole: treebanks, automatic morphological analysis and lexicons.

Treebanks

A treebank can essentially be defined as a "database of sentences which are annotated with syntactic information, often in the form of a tree."[89] Treebanks can be either manually or automatically constructed and they are used to support a variety of computational tasks such as in corpus linguistics, studying syntactic features in computational linguistics and for training and testing parsers. There has been a large growth in the number of historical treebanks in recent years including treebanks in Greek and Latin. Currently there are two major treebank projects for Latin, the Perseus Latin Dependency Treebank (classical Latin) and the Index Thomisticus Treebank (Medieval Latin) and one in Greek, the Perseus Ancient Greek Dependency Treebank (AGDT).[90] This section will briefly look at these treebanks and their uses within classical scholarship.

The Latin Dependency Treebank is a 53,143 word collection of syntactically parsed Latin sentences and it currently stands at version 1.5 with excerpts from eight authors. As Latin is a heavily inflected language with a great degree of variability in its word order, the annotation style of the Latin Dependency Treebank was based off of the Prague Dependency Treebank (PDT), which was then tailored for Latin using the grammar of Pinkster (Bamman and Crane 2006). According to Bamman and Crane (2006) there are a variety of potential uses for a Latin treebank, including "the potential to be used as a knowledge source in a number of traditional lines of inquiry, including rhetoric, lexicography, philology and historical linguistics." In their initial research they explored using the Latin Dependency Treebank to detail the use of Latin rhetorical devices and to quantify the change over time in Latin from a SOV word order to a SVO order. Later research with the use of the Latin Dependency Treebank made use of the resources within the Perseus Digital Library to provide advanced reading support and to provide more sophisticated levels of lemmatized and morpho-syntactic searching (Bamman and Crane 2007).

The Index Thomisticus Treebank (IT-Treebank) is an ongoing project that will include all of the works of Thomas Aquinas as well as 61 authors related to him and will ultimately include 179 texts and 11,000,000 tokens. According to its website, IT-Treebank is "presently composed of 82,141 tokens, for a total of 3,714 syntactically parsed sentences excerpted from *Scriptum super Sententiis Magistri Petri Lombardi*, *Summa contra Gentiles* and *Summa Theologiae*." Their most recent work has explored the development of a valency lexicon, and the authors argue that although many classical languages projects exist, few have annotated texts above the morphological level (McGillivray and Passarotti 2009). Nonetheless, the authors insist that: "nowadays it is possible and indeed necessary to match lexicons with data from (annotated) corpora, and vice versa. This requires the scholars to exploit the vast amount of textual data from classical languages already available in digital format... and particularly those annotated at the highest levels."

Rather than develop their own individual annotation standards, these two Latin treebank projects worked together to develop a common standard set of guidelines that they have published online.[91] This provides an important example of the need for different projects with similar goals to not only collaborate together but also to make the results of that collaboration available for others. Another important collaborative feature of these treebanks is that, particularly in the case of the Perseus Latin Dependency Treebank, a large number of graduate and undergraduate students have contributed to this knowledge base.

Other work in terms of collaborative treebanks has been conducted by the Perseus Digital Library, which has created the Ancient Greek Dependency Treebank (AGDT). The AGDT, currently in version 1.1, is a "192,204-word collection of syntactically parsed Greek sentences" from Hesiod, Homer and Aeschylus. The development of the AGDT, however, has also focused on a new model of treebanking, that of the creation of scholarly

[89] http://en.wiktionary.org/wiki/treebank
[90] For more on the Perseus Latin Dependency Treebank and the AGDT (as well as to download them), see, http://nlp.perseus.tufts.edu/syntax/treebank/, and for the Index Thomisticus Treebank, http://itreebank.marginalia.it/
[91] For the most recent version of the guidelines, see http://hdl.handle.net/10427/42683, and for more on the collaboration see (Bamman, Passarotti and Crane 2008).

treebanks (Bamman, Mambrini and Crane 2009). While traditional linguistic annotation projects have focused on creating the single best annotation (often enforcing inter-annotator agreement), such a model is poor fit when the object of annotation itself is an object of intense scholarly debate:

> In these cases we must provide a means for encoding multiple annotations for a text and allowing scholars who disagree with a specific annotation to encode their disagreement in a quantifiable form. For historical texts especially, scholarly disagreement can be found not only on the level of the correct syntactic parse, but also on the form of the text itself (Bamman, Mambrini and Crane 2009).

The text of Aeschylus serves as a useful example they argue, for many scholars would disagree not just on how a text had been annotated but on the reconstructed text (i.e. the edition that was chosen for annotated) that was itself used. The authors argue that the process of creating scholarly treebanks is similar to that of creating critical editions:

> As the product of scholarly labor, a critical edition displays the text as it is reconstructed by an editor; it is thus an interpretative hypothesis whose foundations lie on the methods of textual criticism. A scholarly treebank may be defined by analogy as a syntactically annotated corpus that again reflects an interpretation of a single scholar, based not only on the scholar's philological acumen but also on an inevitable degree of personal taste and opinions that are culturally and historically determined. A scholarly treebank thus distances itself from the notion that linguistic annotations can be absolute; when dealing with non-native historical languages especially, a syntactic interpretation of a sentence is always the interpretation of an individual and therefore subject to debate (Bamman, Mambrini and Crane 2009).

In order to address this issue, the AGDT thus focused on creating a model that specifically allowed for assigning *authorship* to all interpretative annotations. By doing so, the authors hoped to achieve two goals, first, by publicly releasing the data with citable ownership they wanted to provide a core data set around which scholars could add their own annotations, and second, they hoped that by publicly acknowledging the creators of annotations they could promote the idea of *scholarly treebanking* as an act of scholarly *publication* that is similar in form to publishing a critical edition or commentary. Additionally, they also hoped that this annotation model that gave individual recognition to student contributions to a treebank would serve as one possible model of incorporating undergraduate research into classical teaching. Indeed, many of these issues will be revisited through this review, specifically the need for digital infrastructure to support multiple annotations of different scholars (regarding opinion, certainty, etc.), the ability to show that the creation of digital objects is in itself an act of interpretative scholarship, the importance of attributable and citable scholarship, and the need to support new models of collaboration.

Morphological Analysis

Some of the challenges of automatic morphological processing have already been previously discussed for Sanskrit (Huet 2004) and Sumerian (Tablan et al. 2006), and this subsection will focus briefly on some recent research work in Greek and Latin.

Classical Greek is a highly inflected language and this poses challenges for both students and scholars as detailed by John Lee:

> Indeed, a staple exercise for students of ancient Greek is to identify the root form of an inflected verb. This skill is essential; without knowing the root form, one cannot understand the meaning of the word, or even look it up in a dictionary. For Classics scholars, these myriad forms also pose formidable challenges. In order to search for occurrences of a word in a corpus, all of its forms must be enumerated, since words do not frequently appear in their root forms. This procedure becomes extremely labor-intensive for small words that overlap with other common words (Lee 2008).

The Greek morphological parser for the Perseus Digital Library (named Morpheus) has been in continuous development since 1990 and was developed by Gregory Crane (Crane 1991). Crane worked with a database of 40,000 stems, 13,000 inflections and 2,500 irregular forms. In 1991, Morpheus had been used to analyze almost 3,000,000 words with texts that ranged in data from the eighth century B.C. until the second century A.D. Since this time, Morpheus has played an integral part of the online Perseus Digital Library, which has expanded to cover over 8,000,000 words in Greek. Crane argued that the parser was developed not just to address problems in Ancient Greek but to serve as just one of many possible approaches to developing morphological tools for ancient languages.

More recent work in automatic morphological analysis of Greek has utilized Morpheus as well as other resources available from Perseus. Dik and Whaling (2009) have discussed their implementation of Greek morphological searching over the Perseus Greek corpus that made use of two disambiguated Greek corpora, the open source part-of speech analyzer TreeTagger,[92] and output from Perseus's Morpheus tool. The main backbone of their implementation is a SQLite database backend containing tokens and parses for the full corpus that connects the three main components: the Perseus XML files with unique token IDs, TreeTagger, "which accepts token sequences from the database and outputs parses and probability weights, which are stored in their own table" and PhiloLogic.[93] According to the Dik and Whaling, their system made use of PhiloLogic, because

….it serves as a highly efficient search and retrieval front end, by indexing the augmented XML files as well as the contents of the linked SQLite tables. PhiloLogic's highly optimized index architecture allows near-instantaneous results on complex inquiries such as 'any infinitive forms within 25 words of (dative singulars of) lemma X and string Y', which would be a challenge for typical relational database systems (Dik and Whaling 2009).

The results of their work are available at "Perseus Under PhiloLogic,"[94] a website that supports morphological searching of both the Latin and Greek texts of Perseus. Although Dik and Whaling noted that they were continuing to explore the possibilities of natural language searching against the Greek corpus in place of the very technical ways supported through PhiloLogic, their system nonetheless supports full morphological searching, string searching, lemmatized searching, and these features have been integrated into a reading and browsing environment for the texts.

Some other recent research has focused on the use of machine learning and large unlabeled corpora to perform automatic morphological analysis on classical Greek. Lee (2008) has developed an analyzer of Ancient Greek that "infers the root form of a word" and has made two major innovations over previous systems:

First, *it utilizes a nearest neighbor framework* that requires no hand-crafted rules, and provides analogies to facilitate learning. Second, and perhaps more significantly, *it exploits a large, unlabelled corpus to improve the prediction of novel roots* (Lee 2008).

Lee observed that many students of Ancient Greek memorized "paradigmatic" verbs that could be used as analogies to identify the roots of unseen verbs. From this insight, Lee utilized a nearest neighbor machine-learning framework to model this process. When given a word in an inflected form, the algorithm searched for the root form among its "neighbors" by making substitutions to its prefix and suffix. Valid substitutions are harvested from pairs of inflected and root forms in a training set of data and these pairs are then used to serve as "analogies to reinforce learning." Nonetheless, Ancient Greek still posed some challenges that complicated a minimally supervised approach. Lee explained that heavily inflected languages such as Greek suffer from "data sparseness" since many inflected forms appear at most a few times and many root forms may not appear at all in a corpus. As a rule-based system, Morpheus needed *a priori* knowledge of possible stems and affixes, all of which had to be crafted by hand. In order to provide a more scalable approach, Lee used a data-driven approach that automatically determined stems and affixes from training data (morphology data for the Greek Septuagint from the University of Pennsylvania) and then used the TLG as a source of unlabelled data to guide prediction of novel roots.

While Lee made use of machine learning and unlabeled corpora, Tambouratzis (2008) automated the morphological segmentation of Greek by "coupling an iterative pattern-recognition algorithm with a modest amount of linguistic knowledge, expressed via a set of interactions associated with weights." He used a "ant colony optimization (ACO) metaheuristic" to automatically determine optimal weight values and found that in several cases the automatic system provided better results than when weights had been manually determined by scholars. In contrast to Lee, only a subset of the TLG was used for training data, in this case the speeches of several Greek orators.

---

[92] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[93] PhiloLogic (http://www.lib.uchicago.edu/efts/ARTFL/philologic/) is a software tool that has been developed by the ARTFL project at the University of Chicago, and in "its simplest form serves as a document retrieval or look up mechanism whereby users can search a relational database to retrieve given documents and, in some implementations, portions of texts such as acts, scenes, articles, or head-words."

[94] http://perseus.uchicago.edu/

In addition to the work done by Dik and Whaling for "Perseus Under PhiloLogic", other research into automatic morphological analysis of Latin has been conducted by (Finkel and Stump 2009). These authors reported on computational experiments to generate the morphology of Latin verbs.

Lexicons

Lexicons are important reference tools and have long played an important role in classical scholarship and particularly in the study of historical languages.[95] As was previously noted, the lack of a computational lexicon for Sanskrit is considered to be a major research challenge. This section will explore some important lexicons for classical languages and then explore new roles these traditional reference works could play in a digital environment.

The Comprehensive Aramaic Lexicon[96] (CAL) is a resource that hopes to serve as a "new dictionary of the Aramaic language." Aramaic is a Semitic language known from the beginning of human history and there are numerous inscriptions, papyri as well as Biblical and other religious texts written in this language. This project, currently in preparation by an international team of scholars, is based at Hebrew Union College in Cincinnati. Their goal is to provide a new comprehensive lexicon that will take all of ancient Aramaic into account, be based on a compilation of all Aramaic literature and include extensive references to modern scholarly literature. Although a printed publication is ultimately planned, various databases of textual, lexical and bibliographical information will be available online. Currently a limited version of the lexicon and the bibliographical archives can be searched online.

The Thesaurae Linguae Latinae (TLL)[97] is working on producing "the first comprehensive scholarly dictionary of ancient Latin from the earliest times down to AD 600." This work is based on an archive of about 10 million slips and takes into account all surviving texts. While in older texts there is a slip for every word occurrence, later texts are generally covered by a selection of "lexicographically relevant examples." As Hillen (2007) has explained, from about Apuleius until 600 A.D., textual sources have been excerpted marking noteworthy usages rather than every usage of the word (with the exceptions of major texts by Augustine, Tertullian and Commodian). In order to speed up work, Hillen observed that methods must be found to reduce the number of slips that will be given comprehensive treatment since they outnumber the excerpted material (4.5 to 1) and that emphasis must be given to texts that did not conform to grammatical or stylistic norms. In terms of new digital collections, Hillen saw them as having some value for the work of the TLL:

> The main value of digital databanks for the work of the Thesaurus cannot, therefore, be a systematic increase in the raw material. Rather, they are useful in three specific areas: reproduction, checking, and regulated expansion of our sources (Hillen 2007).

By the end of 2009, the project had reached the end of the letter P, and approximately two thirds of the work has been completed. The TLL has been issued in print version and also has an electronic version that is available by subscription from DeGruyer/Saur.[98]

Similar to the TLL's plan of controlled expansion through documenting unusual or noteworthy usages of words, the "Poorly Attested Words in Greek" (PAWAG)[99] project based at the University of Genoa is setting up an electronic dictionary that "gathers together words of Ancient Greek that are either only scantily attested (i.e. with one or few occurrences), inadequately (i.e. characterized by some sort of uncertainty) or in any case problematically, both from a formal and semantic point of view." This database is intended to serve as a supplement to traditional dictionaries that cannot pay sufficient attention to the issue of poorly attested words. There are currently 1548 headwords in this database that can be searched with a string in either Greek or Latin.

---

[95] This section will focus on larger projects that plan to create online or digital lexicons in addition to printed ones, but there are also a number of lexicons for classical languages that have been placed online as PDFs or in other static formats such as the *Chicago Demotic Dictionary* (http://oi.uchicago.edu/research/projects/dem/), while some other projects have scanned historical dictionaries and provided online searching capabilities, such as *Sanskrit, Tamil and Pahlavi Dictionaries*, http://webapps.uni-koeln.de/tamil/

[96] http://cal1.cn.huc.edu/

[97] http://www.thesaurus.badw.de/english/index.htm

[98] http://www.degruyter.de/cont/fb/at/detail.cfm?id=IS-9783110229561-1

[99] http://www.aristarchus.unige.it/pawag/index.php

A more large-scale endeavor is the Greek Lexicon project [100] that is being overseen by the Faculty of Classics at Cambridge University, which plans to release a new Ancient Greek-English lexicon of intermediate size that will take into account the most recent scholarship, replace archaic terminology with up-to-date English, and both re-examine original source material and also add new material that has been discovered since the end of the 19th century. This project has adopted a semantic method of organizing articles and plans to publish a print edition through Cambridge University press as well as an online version in the Perseus Digital Library. Fraser (2008) has provided more information about the ongoing creation of this lexicon and the challenges that this new semantic organization created. In addition to making use of the Perseus Morpheus database, they developed an additional resource:

….because we can predict every word-search that we will eventually want to perform, a program was designed to conduct these searches in advance. Our corpus of texts has been entirely pre-searched for each lemma-form, and the results archived in static HTML (Hypertext Mark-up Language) pages. This constitutes a digital archive of lexicographic 'slips', providing the dictionary writers with immediate access to the searches, and also enabling the citations and their contexts to be archived in a generic format that is not tied to any particular operating system or database program (Fraser 2008).

This digital archive of Greek lemma searches has helped to greatly speed up the process of writing entries. Interestingly, as this lexicon has been designed for students, Fraser noted that it gives fewer Greek quotations and more space to semantic description. Citations have also been restricted to a canon of seventy authors with no examples taken from fragmentary authors or inscriptions. Fraser also reported that while dictionary entries are stored in XML, the project created a new DTD for their system based on a "provisional entry structure."

The Perseus Digital Library already contains digital versions of lexicons for some individual authors [101] as well as several major classical lexicons, such as the Lewis & Short Latin Dictionary, [102] and the Liddell Scott Johnson Greek English Lexicon (LSJ). [103] The lexicons that are a part of Perseus, however, differ from the previous projects described above. Instead of designing lexicons for both print and electronic distribution, the lexicons and indeed all reference works that are part of Perseus have been created from the start to serve as both hyperlinked tools in an integrated Greek and Latin reading environment and knowledge sources that can be mined to support a variety of automated processes. [104]

In addition to turning "traditional" printed lexicons into dynamic reference works, current research at Perseus is exploring how to create a new kind of "dynamic lexicon" that is generated not from just one printed text but from all the texts in a digital library (Bamman and Crane 2008). They first used the large aligned parallel corpus of English and Latin in Perseus to induce a word sense inventory and determined how often certain definitions of a word were actually manifested, while also using the context surrounding words to determine which definitions were used in a given instance. The treebank was then used to train an automatic syntactic parser for the Latin corpus, in particular to extract information about word's sub-categorization frames and selectional preferences. Clustering was then used to establish semantic similarity between words determined by their appearance in similar contexts (Bamman and Crane 2009). This automatically extracted lexical information can then be used in a variety of ways:

A digital library architecture interacts with this knowledge in three ways: first, it lets us further contextualize our source texts for the users of our existing digital library; second, it allows us to present customized reports for word usage according to the metadata associated with the texts from which they're drawn, enabling us to create a dynamic lexicon that not only notes how a word is used in Latin in general, but also in any specific author, genre, or era (or combination of those). And third, it lets us continue to mine more texts for the knowledge they contain as they're added to the library collection, essentially making it an open-ended service (Bamman and Crane 2008).

As one example, Bamman and Crane (2008) traced how the use of the word *libero* changed over time and across genre (e.g. classical authors vs. Church Fathers). Even though the Perseus corpus is somewhat small, they noted

---

[100] http://www.classics.cam.ac.uk/faculty/research_groups_and_societies/greek_lexicon/
[101] For example, Pindar --http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3atext%3a1999.04.0072
[102] http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3atext%3a1999.04.0059
[103] http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3atext%3a1999.04.0057, for more on the development of the LSJ see (Crane 1998) and (Rydberg Cox 2002).
[104] For more on the development of the LSJ and lexicons in Perseus see (Crane 1998) and (Rydberg-Cox 2002), for more on the need to design "dynamic reference works" see (Crane 2005) and (Crane and Jones 2006).

that even more interesting results could be gained by using such techniques with the large corpus of Latin that is growing online, from multiple editions of classical Latin authors to neo-Latin texts. In a larger corpus, a dynamic lexicon could be used to explore how classical Latin authors such as Caesar and Ovid used words differently, or the use of a word could be compared between classical and neo-Latin texts. Another advantage of a dynamic lexicon is that rather than presenting several highly illustrative examples of word usage (as is done with the Cambridge Greek English Lexicon), it can present as many examples are found in the corpus. Finally, the fact that the dynamic lexicon supports the ability to search across Latin and Greeks text using English translations of Greek and Latin words is a "close approximation to real cross-language information retrieval."

Perhaps most importantly, Bamman and Crane argue that their work to create a dynamic lexicon illustrates how even small structured knowledge sources can be used to mine interesting patterns from larger collections:

> The application of structured knowledge to much larger but unstructured collections addresses a gap left by the massive digitization efforts of groups such as Google and the Open Content Alliance (OCA). While these large projects are creating truly million book collections, the services they provide are general (e.g., key term extraction, named entity analysis, related works) and reflect the wide array of texts and languages they contain. By applying the language specific knowledge of experts (as encoded in our treebank), we are able to create more specific services to complement these general ones already in place. In creating a dynamic lexicon built from the intersection of a 3.5 million word corpus and a 30,457 word treebank, we are highlighting the immense role than even very small structured knowledge sources can play (Bamman and Crane 2008).

The authors also observed that since many of the technologies used to build the lexicon such as word sense disambiguation and syntactic parsing are modular, any separate improvements made to these algorithms could be incorporated back into the lexicon. Similarly, as tagging and parsing accuracy improve with the size of a corpus and as the training corpus of Latin grows in size, so will the treebank. In addition, this work illustrates the importance of how small domain tools might be repurposed to work with larger collections.

Bamman and Crane (2009) have investigated these issues further in their overview of computational linguistics and lexicography. They noted that while the TLG and Perseus provide "dirty results" or the ability to find all the instances of a lemma in their collections, the TLL gives a smaller subset of impeccably precise results. Bamman and Crane argued that in the future, a combination of these two approaches will be necessary, and lexicography will need to both utilize machine learning techniques that learn from large textual collections and utilize the knowledge and labor invested in handcrafted lexica to help such techniques learn. The authors also noted that new lexicons built for a classical cyberinfrastructure would need to support new levels of research:

> Manual lexicography has produced fantastic results for Classical languages, but as we design a cyberinfrastructure for Classics in the future, our aim must be to build a scaffolding that is essentially enabling: it must not only make historical languages more accessible on a functional level, but intellectually as well; it must give students the resources they need to understand a text while also providing scholars the tools to interact with it in whatever ways they see fit (Bamman and Crane 2009).

As this research indicates, lexicons and other traditional reference tools will need to be redesigned as knowledge sources that can be used not just by scholars but by students as well.

Canonical Text Services, Citation Detection, Citation Linking

Digital libraries of classics typically contain both primary and secondary materials (commentaries, dictionaries, lexicons, etc.) Many of these secondary materials as well as journal articles in JSTOR [105] and historical books in Google Books and the Internet Archive will contain a fair amount of latent semantic information in them including references to canonical texts (typically primary sources), historical persons and place names as well as a variety of other information.

In order to effectively link to primary sources, however, these sources must not only be available online but also be structured in a uniform or at least machine actionable way. One proposed solution to this problem is the Canonical Text Services (CTS) protocol. [106] Developed by Neel Smith and Christopher Blackwell, "the Canonical Text Services (CTS) are part of the CITE architecture" and the "specification defines a network service for identifying texts and for retrieving fragments of texts by canonical reference expressed as CTS

[105] http://www.jstor.org/
[106] http://chs75.chs.harvard.edu/projects/diginc/techpub/cts

URNs." [107] Canonical references have been defined as "references to discrete corpora of ancient texts that are written by scholars in a canonical citation format" (Romanello 2008), so for example Hom. *Il*. typically refers to Homer's *Iliad*. Previously known as the Classical Text Services protocol, one major function of this protocol "is to define a network service enabling use of a distributed collection of texts according to notions that are traditional among classicists" (Porter et al. 2006).

The CTS protocol is part of a larger CITE architecture that has been designed to encompass collections of structured objects, indexes, texts and extended objects. The CTS is one of three services defined by this architecture, with the other two being Collection Services and a Reference Index Service. [108] While the Collections Service is still being defined and seeks to "provide an abstract interface to sets of similarly structured objects" the more explicitly defined and mature Reference Indexing or RefIndex service "associates a permanent reference (a CTS URN, or a Collection identifier) with either a second permanent reference, or a raw data value." Reference indexing services encompass mappings traditionally called indices, such as a lemmatized index of a text as well as other kinds such as the mapping of a commentary onto relevant parts of the text.

The most thoroughly defined component of the CITE architecture is the CTS specification/ protocol/ service. The CTS protocol extends the hierarchy of the Functional Requirements for Bibliographic Records (FRBR) conceptual model for bibliographical information developed by the International Federation of Library Associations (IFLA). [109] FRBR defines a *work* as a "distinct intellectual or artistic creation," an *expression* as "the intellectual or artistic realization of a work," a *manifestation* as "the physical embodiment of an expression of a work" and an *item* as "a single exemplar of a manifestation" (IFLA 1998). So in other words, Homer's *Iliad* is a work but an English translation by a particular translator is an expression, an 1890 printing of that particular translation by Macmillan is a manifestation, and an individual copy of that printing on the library shelf is an item.

While the FRBR hierarchy includes Works, Expressions, Manifestations and Items, the CTS protocol uses the terms Work, Edition, Translation and Exemplar. As communicated by Porter et al. (2006), CTS extends the FRBR hierarchy upwards by "grouping Works under a notional entity called "TextGroup'" an entity that can refer to authors for literary texts or to corpora such as inscriptions (e.g. "Berlin" for a published corpus of papyri), and also extends it downward to support the "identification and abstraction of citable chunks of text (Homer, Iliad Book 1, Line 123), or ranges of citable chunks (Hom.~ Il. 1.123-2.22)." This both downward and upward extension of FRBR is important, for as Romanello (2008) underscores "it allows one to reach a higher granularity when accessing documents hierarchically and supports the use of a citation scheme referring to each level of the entire document hierarchical structure." Another important feature of CTS listed by Romanello is that it enables the differentiation of "different exemplars" of the same text.

As noted above, citations in CTS are expressed as CTS URNs. CTS URNs "provide the permanent canonical references that Canonical Text Services (CTS) rely on in order to identify or retrieve passages of text. These references are a kind of Uniform Resource Name (URN)." [110] URN's according to RFC2141, "are intended to serve as persistent, location-independent, resource identifiers." Smith (2009) provides extensive information on the syntax of CTS-URNs and their role in the CTS.

The importance of standards such as CTS is also addressed by this same article (Smith 2009). "Source citation is just one part of scholarly publication, and conventions for citing resources digitally must be viewed as part of a larger architectural design," Smith explained, "when the digital library is the global internet, the natural architecture for scholarly publications is a hierarchy of service." In addition to the need for standards and conventions for citing resources or parts of resources, the importance of standard conventions or ontologies for adding semantic encoding for named entities (such as citations) to secondary literature and also to web pages

[107] For an overview of how CTS URNs and the CITE architecture have been used in the Homer Multitext see (Smith 2010).
[108] http://chs75.chs.harvard.edu/projects/diginc/techpub/cite
[109] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records
[110] http://chs75.chs.harvard.edu/projects/diginc/techpub/cts-urn-overview

and then link them to online representations of primary and other sources has been the subject of a series of recent blog posts by Sebastian Heath of the American Numismatics Society (ANS). [111]

In a recent post entitled "RDFa Patterns for Ancient World References," [112] he described his efforts to encode the year, different named entities such as Polemon, an imperial cult, and two text citations within a chosen text. Heath wanted to embed this information into a sample web page using standards such as RDFa [113] in order to make this data "automatically recognizable by third-parties." In order to encode this information he utilized both RDFa and a number of other ontologies with resolvable namespaces (Dbpedia, cito, foaf, frbr, geo, owl, rdfs, skos). Heath listed two references within his sample text, the first to a published inscription and the second to a recently published book. Interestingly, while Heath was able to link to a bibliographic description of the published book within WorldCat, encoding the reference to the published inscription with the CITO ontology was problematic because there was no value for "cites as a primary source" available within this ontology. An additional complication was that Heath simply wanted to cite the individual inscription and there was no way to cite just one inscription or "work" within the larger published volume of inscriptions. "The concept of "Primary Source" and references thereto is important for the Humanities and we need a way of indicating its usage," Heath concluded, "It's also important that I'm referring to the publication of the inscription, not the inscription itself. When a digital surrogate becomes available, I can point to that. In the meantime, a way of standardizing references to parts of a work would be useful."

Other recent research has also examined some potential methods for resolving these issues of semantic encoding and linking. Romanello (2008) proposed the use of microformats [114] and the CTS to provide semantic linking between classics e-journals and the primary sources/canonical texts they referenced. One of the first challenges, however, was simply to detect the canonical references themselves, for as Romanello demonstrated, references to ancient texts were often abridged, the abbreviations used for author and work names varied greatly, only some citations included the editors names, and the reference schemes could also differ (e.g. for Aeschylus *Persae*, variant citations included A. Pers., Aesch. Pers., and Aeschyl. Pers.). For this reason, Romanello et al. (2009a) explored the use of machine learning to extract canonical references to primary classical sources from unstructured texts. Although references to primary sources found within the secondary literature can vary greatly as seen above, they noted that a number of similar patterns could often be detected. They thus trained conditional random fields (CRF) to identify references to primary sources texts within larger unstructured texts. CRF was a particularly suitable algorithm due to its ability to consider a large number of token features when classifying training data as either "citations" or "not citations." Preliminary results on a sample of 24 pages achieved a precision of 81% and a recall of 94.1%. [115]

Even when references are successfully identified, however, the challenges of encoding and linking still remain. Romanello (2008) stated that most references to primary texts within electronic secondary sources were hard linked "through a tightly coupled linking system" and were also rarely encoded in a machine-readable format. Other obstacles to semantic linking included the lack of shared standards or best practices in terms of encoding primary references in most corpora served as XHTML documents and the lack of common protocols to support interoperability among different texts collections that would thus allow the linking of primary and secondary sources. In order to allow as much interoperability as possible, Romanello thus promoted using "a common protocol to access collections of texts and a shared format to encode canonical references within web online resources" (Romanello 2008). The other requirements of a semantic linking system were that it must be open-ended, interoperable and both semantic and language neutral. Language neutral and unique identifiers for

---

[111] http://numismatics.org

[112] http://mediterraneanceramics.blogspot.com/2010/01/rdfa-patterns-for-ancient-world.html

[113] "RDFa is a specification for attributes to express structured data in any markup language" (http://www.w3.org/TR/rdfa-syntax/) see also http://www.w3.org/TR/xhtml-rdfa-primer/

[114] According to the microformats website, "microformats are a set of simple, open data formats built upon existing and widely adopted standards" that have been designed to be both human and machine readable (http://microformats.org/about)

[115] Work by Romanello continues in this area through crefex(Canonical REFerences Extractor- http://code.google.com/p/crefex/) and was presented at the Digital Classicist/ICS Work in Progress Seminar in July of 2010, see Matteo Romanello, "Towards a Tool for the Automatic Extraction of Canonical References." http://www.digitalclassicist.org/wip/wip2010-04mr.pdf

authors and works (such as those of the TLG) were also recommended in order to support cross-linking across languages.

The basic system for semantic linking outlined by Romanello thus made use of the CTS URN scheme, which utilizes the TLG Canon [116] of Greek authors for identifiers, a series of microformats that he specifically developed to embed canonical references in HTML elements, and open protocols such as the CTS text retrieval protocol to retrieve either whole texts or parts of texts in order to support various value added services such as reference indexing. Romanello proposed three microformats for his system: *ctauthor* (references to canonical authors, or statements that can be made machine readable through the CTS URN structure), *ctwork* (references to works without author names), and *ctref*—"a compound microformat to encode a complete canonical reference" that requires the use of ctauthor, ctwork and a range property to specify the text sections that were referred to. While implementation of such microformats encoding and CTS protocols would enable a number of interesting value added services such as semantic linking, granular text retrieval, and cross lingual reference indexing (e.g. find all articles that reference Verg. *Aen*. regardless of the language they are written in) they also require, as Romanello admitted, a high level of participation by both classical e-journals and relevant digital collections in terms of implementing such microformats and CTS protocols, a factor that seems unlikely.

Another approach to the challenge of semantic linking has been introduced by a project between the classics department at Cornell, which hosts *L'Annee philologique*, and Cornell University Library (Ruddy 2009). This project was awarded a Mellon Planning Grant to explore using OpenURL [117] to provide links from canonical citations in *L'Annee philologique* to their full text in both commercial and open access classics digital libraries. OpenURL was chosen since it provides a uniform linking syntax that was system/vendor independent and minimized the cost of creating and maintaining links. Other stated advantages of OpenURL were that it easily allowed both one-to-many linking and "appropriate copy linking," or the ability to link a user to content they are licensed to see. For example, if a user from outside the library community tried to access a restricted resource such as the TLG they could be directed to the Perseus Digital Library instead.

One of the major project tasks therefore was to create a metadata format that could "reliably reference canonical citations" (Ruddy 2009). The encoding of classical author names was particularly problematic since OpenURL metadata presupposes a modern Western name for an author. Ruddy reasoned that any metadata format would have to allow multiple ways of encoding author forms. In terms of citation structure itself, they did not adopt the CTS but instead chose an abstract approach to recognize the typical hierarchical structure of works.

In addition to metadata challenges there were a number of implementation issues as well. With the normal use of OpenURL, the resolution of a link to a resource is left to a user's local link resolver. The type of solution chosen by this project, however, includes providing an extra level of detailed knowledge, and as Ruddy noted, only "uncertain commercial incentive for link resolver vendors." To solve this issue, Ruddy proposed and consequently created a "domain-specific community supported knowledge base" that was ultimately titled the Classical Works Knowledge Base (CWKB). [118] The final prototype solution implemented by the project was to use the CWKB as an intermediate resolver/knowledge base that could augment and normalize metadata values, provide specialized linking information and support access to free resources for users without a local link resolver. The basic user scenario they envisioned involved a user clicking on a canonical text reference in a JSTOR article or *L'Annee* article abstract that was encoded with an OpenURL that would then direct the user to the CWKB first (which would provide a normalized authority form of author and title and provide a list of services for that work), and then to the local link resolver (if one was available) and finally to a HTML page with link options such as the library catalog, interlibrary loan, text in original language, or the text in translation. In addition, if a user didn't have a link resolver the service could direct them to appropriate free resources by redirecting them back to the CWKB. Ruddy argued that this model could have wider applications, since it could be "useful to any discipline that cites works independent of specific editions or translations" and also offered one solution for chaining link resolvers and knowledge bases together to provide enhanced services to users.

[116] http://www.tlg.uci.edu/canon/fontsel
[117] http://www.oclc.org/research/activities/openurl/default.htm
[118] http://www.cwkb.org/

Text Mining, Quotation Detection and Authorship Attribution

A number of potential technologies could benefit both from automatic citation detection and from the broader use of more standardized citation encoding in digital corpora, these include text mining applications such as the study of text reuse as well as quotation detection and authorship attribution. While the research presented in this section made use of various text mining and NLP techniques with unlabeled corpora, digital texts with large numbers of citations either automatically or manually marked up could provide useful training data for this kind of work. Regardless of how the information is detected and extracted, however, the ability to examine text reuse, trace quotations [119] and both analyze individual authors and study different patterns of authorship will be increasingly important services expected by users of not just mass digitization projects but of classical digital libraries as well.

The eAQUA project [120] based in Germany is broadly investigating how text mining technologies might be used in the analysis of classical texts through six specific sub-projects (reconstruction of the lost works of the Atthidographers, text reuse in Plato, papyri classification, extraction of templates for inscriptions, metrical analysis of Plautus, and text completion of fragmentary texts). [121] "The main focus of this project is to break down research questions from the field of Classics in a reuseable format fitting with NLP algorithms," Büchler et al. (2008) submitted, "and to apply this type of approach to the data from the Ancient sources." This approach of first determining how classical scholars actually conduct research and then attempting to match those processes with appropriate algorithms shows the importance of understanding the discipline for which you are designing tools. This point is an essential one that will be seen continuously throughout this review.

The basic vision of eAQUA is to present a unified approach consisting of "Data, Algorithms and Applications," and this project specifically addresses both the development of applications (research questions) and algorithms (NLP, text mining, co-occurrence analysis, clustering, classification). Data or corpora from research partners will be imported through standardized data interfaces into an eAQUA portal that is currently being developed. This same portal will also provide access to all of the structured data that is extracted through a variety of web services that can be used by scholars. [122]

One area of active research that is being conducted by the eAQUA project is the use of citation detection and textual reuse in the TLG corpus to investigate "the reception of Plato as a case study of textual reuse on ancient Greek texts" (Büchler and Geßner 2009). In their work, they first extracted word-by-word citations by combining n-gram overlaps and significant terms for several works of Plato, and secondly they loosened the constraints on syntactic word order to find citations. The authors emphasized that developing appropriate visualization tools is essential to study textual reuse since text mining approaches to corpora typically produce a huge amount of data that simply cannot be explored manually. Their paper thus offers several intriguing visualizations including highlighting the differences in citations to works of Plato across time (from the Neoplatonists to the Middle Platonists). Other work in textual reuse has been conducted by John Lee (2007), who explored sentence alignment in the Synoptic Gospels of the Greek New Testament. Lee pointed out that exploring ancient text reuse is a difficult but important task since ancient authors rarely acknowledged their sources and often quoted from memory or combined multiple sources. "Identifying the sources of ancient texts is useful in many ways," Lee stressed, "It helps establish their relative dates. It traces the evolution of ideas. The material quoted, left out or altered in a composition provides much insight into the agenda of its author" (Lee 2007).

---

[119] Preliminary research on quotation identification and tracking has been reported for Google Books (Schilit and Kolak 2008).

[120] http://www.eaqua.net/en/index.php

[121] The computational challenges of automatic metrical analysis and fragmentary texts have received some research attention. For metrical analysis see (Deufert et al. 2010, Eder 2007, Fusi 2008), and for fragmentary texts see (Berti et al. 2009) and (Romanello et al. 2009b). The use of digital technology for inscriptions and papyri will be covered in their respective sections.

[122] According to the W3C, a web service can be defined as "a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standard." (http://www.w3.org/TR/ws-arch/#whatis)

Authorship attribution, or using manual or automatic techniques to determine the authorship of anonymous texts has been previously explored in classical studies (Rudman 1998) and remains a topic of interest. Forstall and Scheirer (2009) presented new methods for authorship attribution based on sound rather than text to Greek and Latin poets and prose authors:

> We present the functional n-gram as a feature well-suited to the analysis of poetry and other sound-sensitive material, working toward a stylistics based on sound rather than text. Using Support Vector Machines (SVM) for text classification, we extend the expression of our results from a single marginal distance or a binary yes/no decision to a more flexible receiver-operator characteristic curve. We apply the same feature methodology to Principle Component Analysis (PCA) in order to validate PCA and to explore its expressive potential (Forstall and Scheirer 2009).

The authors discovered that sounds tested with SVMs produced results that performed at least as well if not better than, function-words in every experiment performed, and thus "concluded that sound can be captured and used effectively as a feature for attributing authorship to a variety of literary texts." Forstall and Scheirer also reported some interesting initial results in exploring the Homeric poems, including testing the argument that this poetry was composed without aid of writing, an issue explored at length by the Homeric Multitext Project. "When the works of Thucydides, a literate prose historian, were projected using the principal components derived from Homer, Thucydides' work not only clustered together but had a much smaller radius than either of the Homeric poems," Forstall and Scheirer contended, "This result agrees with philological arguments for the Homer's works having been produced by a wholly different, oral mode of composition." The work of Forstall and Scheirer is just one example of many among digital classics projects of how computer science methodologies can shed "new light" on old questions.

The Perseus Digital Library has conducted some of its own experiments in automatic quotation identification. Ernst-Gerlach and Crane (2008) introduced an algorithm for the automatic analysis of citations but found that they needed to first manually analyze the structure of quotations in three different reference works of Latin texts to determine text quotation alternation patterns. Their experience confirmed Lee's earlier point that text reuse is rarely word for word, though in this case, it was the quotation practices of nineteenth century reference works that proved problematic rather than ancient authors:

> Quotations are, in practice, often not exact. In some cases, our quotations are based on different editions of a text than those to which we have electronic access and we find occasional variations that reflect different versions of the text. We also found, however, that some quotations – especially in reference works such as lexica and grammars – deliberately modify the quoted text – the goal in such cases is not to replicate the original text but to illustrate a point about lexicography, grammar, or some other topic (Ernst-Gerlach and Crane 2008).

This manual analysis provided a classification of the different types of text variation including regular text differences, irregular text differences, word omission, text insertion and word substitution. The algorithm that was ultimately developed has not as yet been put into the production Perseus Digital Library.

Other research by Perseus has also explored automatic citation and quotation identification by utilizing quotation indices or "indices scriptorum" that typically listed all of the authors quoted within a classical text and were manually created by editors for critical editions of classical texts. Fuzzy parsing techniques were applied to the OCR transcription of one such index from the *Deipnosophistae* of Athenaeus in order to then automatically mark up all the quotations found within the index in a digital version of the text (Romanello et al. 2009c).

The Disciplines and Technologies of Digital Classics

This section will explore a variety of important sub-disciplines or related disciplines of digital classics with an overview of some important projects and relevant literature in each. These overviews are by no-means exhaustive and sought to identify the major projects in each field that illustrate the major challenges faced.

Ancient History

In many ways the study of ancient history is less a sub-discipline of classical studies than an overarching field that makes uses all of the sources combined that are studied intensively in each of the other sub-disciplines, so various aspects of this topic are covered in many of the different sub-discipline sections rather than exclusively

here. The popularity of this topic is evidenced by innumerable academic and enthusiast websites on the history of Greece, Rome and the Ancient Near East. One of the larger enthusiast websites is Attalus.org [123] that provides detailed "lists of events and sources for the history of the Hellenistic world and the Roman republic" and also includes translations of many of the relevant sources such as Livy and Tacitus. The Livius [124] website managed by Dutch historian Jona Lendering offers a search engine to a large number of online articles on Roman and Greek history.

The subject of ancient history makes up a large component of many digital classics projects, albeit not necessarily as a specific focus, rather the sources provided (whether primary texts or documentary sources such as inscriptions, papyri or coins) support the study of ancient history. As one report recently noted, digital archives are providing access to all of these materials at a rapid rate:

> Scholars of ancient history study particular documentary remains (such as inscriptions, papyri, ancient maps, classical literature and drama, and art and architecture) to examine early societies. These materials may be excavated from multiple archaeological sites, and are generally found in archives. While documentary editions have traditionally provided scholars with wider access to archival sources, the growth of digital archives is perceived as a great boon. On the one hand, digital archives are allowing scholars to search, study, and make connections between more archival materials. On the other hand, different archival materials are being digitized at various rates, and looking at a digital surrogate may not replace the value of seeing the physical artifact. The "next generation" of digital archives, according to some scholars, will integrate archival materials with computational linguistics and other text-mining capabilities (Harley et al. 2010, pg. 118).

These themes of "next generation" digital archives and the need to combine them with sophisticated language technologies will be revisited in later sections of this report.

To conclude this section we will briefly cover one major technology project that seeks to support the encoding of historical information wherever it is found, namely the HEML or the Historical Event and Markup Linking Project (Robertson 2009), which seeks to provide markup standards and tools with which to encode historical information on the Web. In a state of continuous evolution since 2001, HEML now has an RDF data model that allows it to represent nested events and relations of causality between events. The HEML data format supports collections of events where each is tagged with heml:Event, which at their simplest are bound to machine readable spans of time and references to evidence. Other important features of the model include the assignment of Uniform Resource Identifiers (URI) [125] to all individual entities and the utilization of an evidence element:

> Persons, roles, locations, and keywords are assigned mandatory URIs so that they may be referred to in multiple events. Finally, one or more heml:Evidence elements must be attributed to each event, and within these there is a means by which different editions and linguistic representations of the same text may be grouped together for the researcher's benefit (Robertson 2009).

These encoding choices illustrate the importance of using unique URIs to identify specific entities not only so they can be referred to in multiple encoded historical events but also so they can be reused as "linked data" [126] by other applications. In addition, the ability to link encoded events to attestations in primary texts is also of critical importance. One criticism often made of HEML Robertson stated is that "it is not possible to encode the variations in opinions regarding historical events." The ability to encode multiple scholarly opinions and to indicate the uncertainty of knowledge regarding dates or other information are both important features of any markup for historical texts. Robertson does argue, however, that URIs could eventually be created for scholars, and specific encoded arguments could be linked to those URIs.

---

[123] http://www.attalus.org/
[124] http://www.livius.org/
[125] A URI has been defined as a "compact sequence of characters that identifies an abstract or physical resource." For more on their syntax and architecture, see http://labs.apache.org/webarch/uri/rfc/rfc3986.html#overview
[126] Tim Berners-Lee has described linked data as essential to the creation of the Semantic Web, and the creation of linked data must follow four essential rules: 1. "uses URIs as names for things" 2. "Use HTTP URIs so that people can look up those names" 3. Use useful standards such as RDF and SPARQL so that when someone looks up a URI it provides useful information 4. Include links to other URIs to support further discovery. http://www.w3.org/DesignIssues/LinkedData.html

Classical Archaeology

Overview

Archaeology is the study of the material remains and environmental effects of human behavior throughout prehistory to the modern era. Scholarship in archaeology is divided into a large number of subdisciplines, many defined geographically (e.g., North America, Egypt, Near East, Oceania) and/or by time period (e.g., Paleolithic, Neolithic, Classical). A moderately sized field, archaeology overlaps with a range of other scholarly disciplines, including biological anthropology, ethnobotany, paleozoology, geology, and classics (in particular, palaeography, philology, papyrology, epigraphy, numismatics, history of the ancient world, Hellenic literature, and art and architectural history) (Harley et al. 2010, pg. 30).

As illustrated by this definition, archaeology is a complex and interdisciplinary field with many of its own specializations but that is also closely related to classics. The innovative use of digital technology in archaeology has a history that is at least three decades old. Numerous conferences contains papers involving the use of 3d-visualizations, digital reconstructions and electronic publication in archaeology—including Computer Applications and Quantitative Methods in Archaeology (CAA), the Virtual Systems and Multimedia Society (VSMM), Visual Analytics Science and Technology Symposium (VAST), and the International Committee for Documentation of Cultural Heritage (CIPA). In addition, the importance of both cyberinfrastructure and digital preservation for archaeology has been addressed by a number of recent projects as well as longer standing organizations. This section will look at several major digital classical archaeology projects and provide a brief overview of some of the literature on this vast topic.

Electronic Publishing and Traditional Publishing

One of the oldest e-journals in archaeology is *Internet Archaeology*, [127] a peer reviewed journal that was established in 1996 and sought to make full use of electronic publishing. There are also some other interesting examples of electronic publication in archaeology including FastiOnline. [128] Nonetheless, a major new report from the Center for Studies in Higher Education (CSHE) at the University of California-Berkeley [129] that investigated the potential of digital scholarship and scholarly communication across a number of disciplines, including archaeology, by interviewing scholars at elite institutions (Harley et al. 2010), found that most archaeologists were still very distrustful of electronic publication, peer reviewed or not.

This same report surveyed how digital technology was affecting the nature of publishing, tenure and scholarship in archaeology. Traditional publishing of archaeological scholarship is typically done through monographs and less frequently through journal articles although in more technical fields such as epigraphy or papyrology, journal articles are more likely to be the standard. The complicated nature of archaeological data, with its extensive use of images and other multimedia in addition to the limitations of print publishing, however, has led many projects to create complex archaeological websites or to pursue more sophisticated digital publishing options. Two interesting examples include Ostia: Harbour City of Ancient Rome [130] and the Pylos Regional Archaeological Project. [131]

Despite this growing practice, the CSHE researchers documented a fairly strong resistance to the consideration of digital publications for tenure dossiers, this was largely due to the lack of precedent, a general uncertainty regarding how to peer review electronic publications, and a belief that such projects often failed to make scholarly arguments. A quote from one scholar demonstrates this general view:

> But I would say the test for me is not do you have computer wizards doing classics—the answer is yes—but instead, are there works in classics that have come up, which are excellent particularly because of their technological connections? I don't know of one. So I think that there is a basic mistrust toward the digital medium in academia, and particularly with regard to tenure and promotion. And I think that to some extent it's justified….At best, the digital medium produces data that are structured sometimes very well and with a great deal

[127] http://intarch.ac.uk/
[128] http://www.fastionline.org/index.php
[129] http://cshe.berkeley.edu/
[130] http://www.ostia-antica.org/
[131] http://classics.uc.edu/PRAP/

of interactive opportunities, search capabilities, and whatnot. But a website does not really develop an argument, and what we expect of a scholar, young or old, is to be able to develop an argument (Harley et al. 2010, pg. 38).

The authors of this report also noted that one department head suggested that digital publications should be represented in a "new category" between service and teaching. Many archaeologists that were interviewed also did not seem to realize that a number of online journals such as *Internet Archaeology* are peer-reviewed. While traditional publishing largely remains the rule, the Center for Hellenic Studies (CHS) [132] has founded its own digital publishing program including journals such as Classics@ and has also put electronic books on its website. [133] On the other hand, the American Philological Association (APA) and Archaeological Institute of America (AIA) released a report on electronic publications in 2007 where the most "revolutionary" move proposed was for the APA to "explore a new digitally-distributed series of APA monographs" (APA/AIA 2007).

Not all scholars interviewed by the CSHE were pessimistic about the potential of electronic publishing, however, and some believed the potential was largely dependent on the discipline. One scholar interviewed felt that it was reasonable that within 10 years "most of the scholarly life cycle in papyrology" would be integrated and accessible in the same technological platform. This same scholar was also very hopeful regarding the digital potential for epigraphy and the development of a comprehensive digital environment for Greek but also argued that far more work needed to be done for Latin. A final insight offered by this same individual was that "these digital projects make it possible for the rest of us to get access to the material that makes the books possible" (Harley et al. 2010, pg. 64). Increasing digital access to archaeological data according to this scholar not only supported digital scholarship but had great benefits for traditional monograph publishing as well.

Despite many traditional scholars reluctance to evaluate web-based publications, the CSHE report maintained that new models of online publication offer not just better ways of integrating all types of data from site reports and databases to videos and digital reconstructions, but can provide almost immediate access to data as it is discovered. Even more importantly, "these initiatives are truly data-driven, collaborative, and require a shift in the traditional thinking of the "monograph as the final publication on a site" (Harley et al. 2010). Online publication thus offers new opportunities for data driven scholarship, collaboration, and almost real-time updating. The report authors also assert that an additional advantage of the dynamic nature of online publication is that scholarly arguments can evolve as more data are published and this process can be made much more transparent to the reader. Meckseper and Warwick (2003) make a similar point that electronic publication can help archaeology as a discipline move towards a better integration of data and published interpretations. They reflected that the practice of treating excavation reports as sole data archives had come under heavy criticism by the 1980s, as more archaeologists came to realize "the distinction between data and interpretation was often not as easy to maintain as previously assumed" (Meckseper and Warwick, 2003). The ability to represent uncertainty regarding individual scholarly interpretations or statements is one of the reasons the authors chose to use TEI to encode archaeological reports.

Stuart Dunn has also discussed how electronic publication in archaeology has both a number of benefits and difficulties. The deposit of archaeological data into digital repositories or virtual research environments (VREs), while still preserving copyright and intellectual property, is a difficult but ultimately worthwhile goal he contends. Nonetheless, the potential of linking, for example, published articles to the data they reference, also raises questions of data accuracy, controlled access, security and transparency:

Where a discussion in a published article focuses on a particular set of primary data, there is a clear logic to deploying VRE tools, where available, to make that data available alongside the discussion. However, in such situations it is incumbent upon the VRE to ensure that those data are trustworthy, or, if they are not (or might not be), to provide transparent documentation about the process(es) of analysis and manipulation via which they have come to support the published discussion….the term "research" in Virtual Research Environment implies that the outputs meet "conventional" standards of peer review and evaluation (Dunn 2009).

[132] The CHS is a classical research institute that is affiliated with Harvard University and is based in Washington, D.C. Founded in 1962 through an independent endowment made "exclusively for the establishment of an educational center in the field of Hellenic Studies designed to re-discover the humanism of the Hellenic Greeks" the CHS hosts a number of innovative digital projects (such as the Homer Multitext) and is also responsible for a number of different online publications.

[133] http://chs.harvard.edu/wa/pageR?tn=Publications&bdc=12&mn=0

Although developing new models of peer review and authentication for digital publication will not be easy challenges to meet, Dunn rightly concludes that this does not make them not worth working towards.

Data Creation, Data Sharing, Data Preservation

Archaeological research provides a wealth of data of greatly different types as the CSHE report summarizes:

> Archaeological research is somewhat exceptional among its humanistic neighbors in its reliance on time- and location-specific data, abundant use of images, and dependence on complex interdisciplinary teams of scholars and specialists, who work on both site excavation and complex lab-based data analysis. Teams produce a plethora of data types in archaeology, including three-dimensional artifacts, maps, sketches, moving and still images, flora and faunal assemblages, geological samples, virtual reconstructions, and field notes. (Harley et al. 2010, pg. 30-31)

These greatly varying kinds of data make the development of any standards for data recording, sharing and preservation a considerable undertaking.

In terms of data sharing, the CSHE report suggested that most archaeological scholars shared ideas through informal networks, email and small meetings, but tended to keep all data and work-in progress to themselves until formal publication. A variety of factors influenced these decisions including various stakeholder interests, fear of data being "poached," the sensitivity of some archaeological sites and the "messiness of the data." On the other hand, papyrologists tended to work together, a factor that shall be discussed in greater detail later. While some scholars were familiar with working papers sites in classics such as the Princeton/Stanford Working Papers in Classics (PSWPC) [134] and the Classics Research Network (CRN), [135] there was not a great deal of interest in whether such a working paper site should be created for archaeology. One scholar interviewed by Harley et al. (2010) succinctly described the problem as being "that archaeology has a culture of ownership of ideas as property, rather than the culture of a gift economy" (Harley et al. 2010, pg. 81). While archaeologists most often collaborated on data collection, they rarely coauthored articles together, a criticism that is also often made of classicists.

Nonetheless, the CSHE report also illustrated that for a growing number of archaeologists the idea of data sharing as data preservation is gaining importance. Since archaeological sites are typically destroyed during an excavation, scholars highlighted the need to be very meticulous in recording all of the necessary information about data and also revealed that a great deal of "dark data' from excavations never makes it to final publication. In other words, the published record and the data from an excavation are the only surviving record of an archaeological site in many cases. Meckseper and Warwick also underscore this fact in their exploration of XML as a means of publishing archaeological excavation reports. "Archaeology is a destructive process: the physical remains in the ground are destroyed through their excavation and lifting of material," Meckseper and Warwick confirmed, "The written record and publication have therefore always been seen as synonymous with the preservation of the archaeological record" (Meckseper and Warwick 2003). Shen et al. (2008) thus also agreed that for this reason the digital recording of data at both the planning and excavation stages are extremely important, for as they noted "unlike many other applications of information systems, it simply is not possible to go back and re-check at a later date."

While many scholars interviewed by the CSHE believed in the need to share more datasets they also lamented the complete lack of standards for sharing and preserving data. The best-known data model for archaeology is Archaeological Markup Language (ArchaeoML), an XML schema for archaeological data (Schloen 2001) that serves as the basis of the XML database of the OCHRE (Online Cultural Heritage Research Environment) project. [136] Based at the University of Chicago, OCHRE is an Internet database system that has been designed to manage cultural heritage information. According to the website, "it is intended for researchers engaged in artifactual and textual studies of various kinds. It is especially suitable (1) for organizing and publishing the results of archaeological excavations and surveys and (2) for preparing and disseminating philological text editions and dictionaries." OCHRE implements a core ontology for cultural heritage information and uses a

[134] http://www.princeton.edu/~pswpc/index.html
[135] http://www.ssrn.com/crn/index.html
[136] http://ochre.lib.uchicago.edu/index.htm

"global schema" to which local schemas of various projects can be mapped in order to facilitate data integration. [137] A number of projects are using OCHRE to present their research including the Chicago Hittite Dictionary [138] and the Persepolis Fortification Archive Project. [139]

The CSHE report also drew attention to the fact that data preservation in archaeology is becoming increasingly problematic due to the need to preserve both analog and digital data. Many scholars that were interviewed wanted more institutional support for storing, migrating and backing up their data. A similar problem is that while archaeological projects funded with public money in the United Kingdom are required to make their data publicly available, publishing mandates differ greatly between U.S. funders.

Open Context, [140] an "open access data publication service for archaeology" created by the Alexandria Archive Institute, is attempting to address some of these issues of data sharing and preservation. In an article by Kansa et al. (2007) that describes Open Context, the authors explicate why data sharing and dissemination are particularly complicated for archaeology:

> Among the primary technical and conceptual issues in sharing field data is the question of how to codify our documentation. Archaeologists generally lack consensus on standards of recording and tend to make their own customized databases to suit the needs of their individual research agendas, theoretical perspectives, and time and budgetary constraints…. Because of this variability, databases need extensive documentation for others to decipher their contents (Kansa et al. 2007).

Consequently the authors propose that just making archaeological datasets available for download will not solve this basic problem and that a better solution is to "serve archaeological databases in dynamic, online websites, thus making content easy to browse and explore." Open Context seeks to make the publishing and dissemination of cultural heritage collections both easier and more affordable. The basic architecture of Open Context is a flexible database that allows researchers to publish structured data, textual narratives, and media on the web using only open source technologies. Open Context supports "publishing, exploring, searching, and analyzing multiple museum collections and field research datasets."

Open Context utilizes only a subset of the "OCHRE data structure (ArchaeoML)" for Kansa states that while OCHRE "provides sophisticated data-management tools targeted for active research projects" the goal of Open Context is "to support streamlined, web-based access and community organization of diverse cultural heritage content" (Kansa et al. 2007). The project ultimately decided to use ArchaeoML due to its flexibility:

> Overly rigid standards may inhibit innovation in research design and poorly accommodate "legacy" datasets…. The flexibility of ArchaeoML enables Open Context to deliver content from many different research projects and collections. A web-based publishing tool called "Penelope" enables individual contributors to upload their own data tables and media files and submit them for review and publication in Open Context. This tool enables web publication of research while ensuring that a project's original recording system and terminology are retained (Kansa et al. 2007).

The ability of standards such as ArchaeoML to provide some basic level of interoperability while also supporting the inclusion of legacy structures is thus an essential feature for any cyberinfrastructure for archaeology.

Another important issue that Open Context also seeks to address is the challenge of open access and copyright. All Open Context contributors retain copyright to their own content, but Kansa et al. also state that they are encouraged to publish their data elsewhere in order to better support dissemination and digital preservation. The authors contend that current copyright laws make digital preservation difficult because permission to copy any data must be granted explicitly by the copyright holder, even if data is copied simply to back it up, thus they require all contributors to use copyright licenses that grant permissions to reproduce content. Kansa et al. also insisted that "copyright will typically apply to most archaeological field data" in order to assuage the concern of most archaeologists that if they place field data online before formally publishing it, that their data will be stolen. This point, however, has been challenged by others, who assert that any field data published online will

---

[137] For a good overview of this system and how it compares to the CIDOC-CRM and to tDAR of Digital Antiquity, please see http://ochre.lib.uchicago.edu/index_files/Page794.htm
[138] http://ochre.lib.uchicago.edu/eCHD/
[139] http://ochre.lib.uchicago.edu/PFA_Online
[140] http://opencontext.org/

enter the public domain immediately. While this legal debate is beyond the scope of this report, Open Context also supports other features such as time-stamping the accession of new collections, clearly identifying authorship, and providing permanent citable URLs in an effort to encourage proper citation and reuse of data. The creators of Open Context ultimately hope that the development of this system will serve as one step in increasing open access in the field of archaeology.

Although collaboration and data sharing are often not the norm, the field of research archaeology in particular requires a great deal of collaboration due to the large number of specialized fields it involves. Another area where the CSHE report listed increasing collaboration was between domain specialists in archaeology and technical experts. At the same time, many archaeologists argued that more domain specialists needed to become technical experts as well in order to design tools for the field that would actually be used:

> Indeed, some scholars who specialize in such areas as "virtual heritage" consider themselves to be "methodologists of archaeological research" or "technological ambassadors," rather than particular experts in a specific period or culture. Moreover, some scholars with dual archaeological and technical expertise may turn to parallel career paths, and may play an important, often central, role in creating the infrastructure for successful scholarship (Harley et al. 2010, pg. 109)

The growing need to master both domain and technical expertise is a theme that shall be seen throughout the overviews of all the digital classical disciplines.

Digital Repositories, Data Integration & Cyberinfrastructure for Archaeology

Arguably one of the best-known repositories for archaeological data is the Archaeology Data Service (ADS) [141] based at the University of York in the United Kingdom. The ADS provides digital archiving services for archaeology projects within the U.K., a searchable catalogue of projects and their data (labeled ArchSearch) and promotes best practices for digitization, preservation and database management in the larger field of archaeology. Once part of the now defunct Arts & Humanities Data Services (AHDS) [142], the ADS receives funding from the Arts and Humanities Research Council (AHRC) but also has a charging policy with set fees for storage and dissemination. [143]

Several recent initiatives have also sought to provide the beginnings of a cyberinfrastructure for archaeology. The Mellon funded Archaeoinformatics [144] was "established as a collaborative organization to design, seek funding for, and direct a set of cyberinfrastructure initiatives for archaeology." This project ran from 2007 to 2008, and according to their project summary, this initiative sought to provide preservation for both archaeological data and metadata and to build a digital archive for data that would provide access to both scholars and the general public. Digital Antiquity, [145] the successor to Archaeoinformatics, is a "collaborative organization devoted to enhancing preservation and access to digital records of archaeological investigation." With funding from Mellon and the National Science Foundation (NSF), this project plans over the next two years to build an "on-line digital repository that is able to provide preservation, discovery, and access for data and documents produced by archaeological projects." Named tDAR (the Digital Archaeological Record") this repository plans to encompass digital data from both ongoing research and legacy archaeological projects, with a focus on American archaeology. After initial funding endings, Digital Antiquity plans to utilize a data curation model:

> ….those responsible for archaeological investigations will pay a fee for the deposit of digital data and documents. Once deposited, access will be freely available over the Internet upon user registration, consent with a use agreement, and with restrictions for sensitive information. The entrepreneurial, service-oriented focus of the enterprise is necessary for Digital Antiquity to become financially self-sustaining in a 4-5 year period. It is planned, at that point, for the organization to transition into an independent not-for-profit entity or be brought under the umbrella of another appropriate not-for-profit (such as a professional society) that can host the repository in the long-term (Digital Antiquity, 2010).

---

[141] http://ads.ahds.ac.uk/
[142] http://ahds.ac.uk/
[143] http://www.ahrc.ac.uk/Pages/default.aspx
[144] http://archaeoinformatics.org/
[145] http://www.tdar.org/confluence/display/DIGITAQ/Home

This project is currently in its early stages but (Elliott 2008) has provided an overview of the technological architecture of the planned tDAR.

Rather than attempting complete integration of all datasets or designing an universal data model for archaeology, Elliott reported that the "semantic demands" of queries are reconciled with the semantic content of available datasets:

tDAR uses a novel strategy of query-driven, ad-hoc data integration in which, *given a query*, the cybertools will identify relevant data sources and perform interactive, on-the-fly metadata matching to align key portions of the data while reasoning with potentially incomplete and inconsistent information (Elliott 2008).

Currently the prototype tDAR allows users to search an initial data archive and to register and upload resources (including databases, text files, and images). While anyone may register to use tDAR, only approved users can upload and add information resources. Information resources can be either public or private so different levels of access control can be supported. tDAR also supports a variety of data formats including text files in ASCII or PDF, JPEG and TIFF images, and databases can be ingested as Access, Excel, or CSV files. All uploaded databases are converted to a standard relational database format that will serve as the long-term format for preservation and updating.

While a number of scholars interviewed by the CSHE had great hopes for projects such as Archaeoinformatics, most still believed that "one of the biggest obstacles is the question of standards for integrating idiosyncratic data sets on a large scale, as well as the difficulties of securing buy-in from other stakeholders of archaeological data" (Harley et al. 2010, pg. 91). Similarly, Stuart Dunn in his overview of developing a specific VRE in archaeology commented that projects were often good at creating infrastructure, but only at the project level. "Because archaeological fieldwork is by definition regional or site-specific, excavation directors generally focus the majority of their efforts at any one time on relatively small-scale data gathering activities," Dunn declared, "This produces bodies of data that might be conceptually comparable, but are not standardized or consistent" (Dunn 2009).

These challenges have also been articulated by earlier research into the cyberinfrastructure needs of archaeology. An article by Snow et al. (2006) listed three particular types of data that are almost impossible to access simultaneously due to lack of cyberinfrastructure for archaeology: databases using different standards for both recording and managing data on different technical platforms, a large volume of "grey literature," and images, maps and photographs found in museum catalogs and both published and unpublished archaeological reports. The authors proposed a cyberinfrastructure architecture for archaeology that included existing digital library middleware (such as Fedora) [146] and content management tools, document and image searching technologies, geographic information systems (GIS), visualization tools, and the use of the Open Access Initiative Protocol for Metadata Harvesting (OAI-PMH) [147] because it provides an "application-independent interoperability framework for metadata harvesting" by repositories. Snow et al. 2006 also reported that such an infrastructure could be built almost entirely from open-source components.

No such infrastructure has yet been built, however, and Snow et al. also identified several critical problems that must be addressed including the lack of any type of standard protocols for recording the greatly varying types of archaeological data and the absence of any standardized tools for access. Nonetheless the authors also suggested that rather than trying to force the use of one data model, semantic mapping tools should be used to map different terminologies or vocabularies in use, while also working within the archaeological community to establish at least minimal shared standards for description. Finally, in order to ensure sustainability, Snow et al. argued that "data collections should be distributed and sharable" and that "digital libraries and associated services should be made available to researchers and organizations to store their own data and mirror data of others."

More recent research by Pettersen et al. (2008) has reached similar conclusions. This article reported on attempts to create an integrated data grid for two archaeological projects in Australia, and stated that:

[146] Fedora is an advanced digital repository platform (http://www.fedora-commons.org/) that is available as open source.
[147] http://www.openarchives.org/pmh/

A continuing problem in the archaeological and cultural heritage industries is a lack of coordinated digital resources and tools to access, analyze and visualize archaeological data for research and publication. A related problem is the absence of persistent archives that focus on the long-term preservation of these data. As a result professionals and researchers are either unaware of the existence of data sets, or aware of them but unable to access them for a particular project (Pettersen et al. 2008).

One potential benefit of a coordinated cyberinfrastructure or integrated digital archive for more archaeological projects as indicated here is that it would allow more researchers to not only find and possibly reuse data but also to use their own tools with that data (such as visualizations). The architecture ultimately chosen by this project was to utilize the Storage Resource Broker (SRB) developed by the San Diego Supercomputing Center. The biggest challenge they found in using the SRB was its lack of an easy to use interface. Their project also encountered various challenges of data capture in the field, data logging, and they criticized the fact, like so many other researchers, that "there is no standardized methodology in archaeology for recording data in a digital format." While Pettersen et al. (2008) are currently exploring the use of ArchaeoML for data integration and portability, they also submitted that the largest obstacle still to be overcome was to create a user-friendly way for archaeologists to interact with the data grid.

A variety of research by the ETANA-DL [148] has also explored the difficulties of integrating archaeological collections. This digital library is part of the larger project ETANA (Electronic Tools and Ancient Near Eastern Archives), [149] which also provides access to ABZU and includes a collection of core texts in the field of Ancient Near East studies. The ETANA-DL utilized the 5S (streams, structures, spaces, scenarios, and societies) framework to integrate several archaeological digital libraries (Shen et al. 2008). They developed a domain metamodel for archaeology in terms of the 5S model and focused particularly on the challenges of digital library integration. The architecture of ETANA-DL consists of a "centralized catalog and partially decentralized union repository." In order to create the centralized union catalog they utilized mapping and harvesting services. ETANA-DL also continues to provide all the services offered by the individual digital libraries they integrated through what they term union services:

Union services are new implementations of all the services supported by member DLs to be integrated. They apply to the union catalog and the union repository that are integrated from member DLs. The union services do not communicate with member DLs directly and thus do not rely on member DLs to provide services (Shen et al. 2008).

The authors stress the importance of providing integrated user services over an integrated digital library, while still developing an architecture that allows the individual libraries to retain their autonomy.

In agreement with Kansa et al. (2007), the creators of ETANA-DL also warned against attempting to create one universal schema for archaeology:

Migration or export of archeological data from one system to another is a monumental task that is aggravated by peculiar data formats and database schemas. Furthermore, archeological data classification depends on a number of vaguely defined qualitative characteristics, which are open to personal interpretation. Different branches of archeology have special methods of classification; progress in digs and new types of excavated finds make it impossible to foresee an ultimate global schema for the description of all excavation data… Accordingly, an ''incremental'' approach is desired for global schema enrichment (Shen et al. 2008).

Instead of using ArchaeoML or a subset, as with Open Context, the creators of ETANA-DL have incrementally created a global schema and support data integration between different digital libraries through the use of "an interactive software tool for database-to-XML generation, schema mapping, and global archive generation" (Vemuri et al. 2006). The three major components to the ETANA-ADD tool are a database to XML converter, a schema mapper, and an OAI-XML data provider tool. The first component DB2XML converts data from custom databases into XML collections. "The end user can open tables corresponding to an artifact, and call for an SQL join operation on them. Each record of the result represents an XML record in the collection, and the structure of the dataset determines the local XML schema," Vemuri et al. (2006) explained, "Based on this principle, the component generates a local XML collection and its XML schema." After a local XML collection is generated, end users interact with a tool called "Schema Mapper" that leads a user through mapping the local XML schema that has been generated for their database into the global XML schema used by ETANA-DL. If a

[148] http://digbase.etana.org:8080/etana/servlet/Start
[149] http://www.etana.org/

particular artifact type or other item isn't available the global XML schema is extended to include it. The final component is an OAI XML Data provider that supports publishing of the new XML collection as an OAI data provider. Thus the ETANA-DL created a system that allowed for an almost lossless conversion of individual databases into their own universal schema and created an integrated archaeological digital library from 3 individual ones that can now be searched seamlessly.

Designing Digital Infrastructures for the Research Methods of Archaeology

The basic stages in archaeological research have been described as discovery, identification and attribution, cross-referencing, interpretation, and publication (Dunn 2009). In addition, most archaeological research begins with site excavation, which produces massive amounts of data. Thus research challenges in archaeology typically include organizing excavation data after the fact and sophisticated means of data analysis including spatial analysis, 3d modeling of sites and artifact imaging. Many scholars interviewed in the CSHE report also argued for a greater need to link the material record that they document so meticulously with the growing textual record, including text collections in digital libraries and printed editions found in mass digitization projects.

Efforts to reintegrate the material and textual records of archaeology were recently explored by the Archaeotools project [150] (Jeffrey et al. 2009a, Jeffrey et al. 2009b). Archaeotools was a major e-Science infrastructure project for archaeology in the UK that sought to create a single faceted browser interface that would integrate access both to the millions of structured database records regarding archaeological sties and monuments found in the ADS with "information extracted from semi-structured grey literature reports, and unstructured antiquarian journal accounts." Archaeotools explored both the use of information extraction techniques with arts and humanities datasets and the automatic creation of metadata for those archaeological reports that had no manually created metadata. Jeffrey et al. (2009b) observed that archaeology has an extensive printed record going back to the 19th century including monographs, journal articles, special society publications and a vast body of grey literature. One unique challenge of much of the antiquarian literature they also noted was the use of non-standard historical place names that made it impossible to automatically integrate this information with modern GIS and mapping technologies. Their project was informed by the results of the Armadillo project, [151] a historical text mining project that used information extraction to identify names of historical persons and places in the Old Bailey Proceedings [152] and then mapped them to a defined ontology.

The Archaeotools project ultimately created a faceted classification and geospatial browser for the ADS database, with the main facets for browsing falling into the categories of: "what," "where," "when" and "media". All facets were populated using existing thesauri that were marked up into XML and then integrated using SKOS. [153] Selected fields were then extracted from the ADS database in MIDAS XML [154] format, converted to RDF XML and then mapped onto the thesauri ontology that was previously created. The project also created an extendable NLP system that automatically extracted metadata from unpublished archaeological reports and legacy historical publications that used a combination of knowledge engineering (KE) and automatic training (AT). The final task was to use the geoXwalk [155] service to recast "historical place names and locations as national grid references."

Despite the growth of projects such as Archaeotools, one scholar interviewed by the CSHE concluded that he had yet to see any revolutionary uses of technology within archaeology. "What I see still is mainly people being able to do much more of what they always were able to do, and do it faster, and in some cases better, with the tools," this scholar observed, "I don't see yet that the technology is fundamentally changing the nature of what people are doing…." (Harley et al. 2010, pg. 120). This argument is seen often in criticism of digital classics

[150] http://ads.ahds.ac.uk/project/archaeotools/
[151] http://www.hrionline.ac.uk/armadillo/objectives.html
[152] http://www.oldbaileyonline.org/
[153] SKOS stands for "Simple Knowledge Organization System" and provides a RDF model for encoding reference tool such as thesauri, taxonomies and classification systems. SKOS is currently under active development as part of the W3C's Semantic Web activity (http://www.w3.org/2004/02/skos/)
[154] http://www.heritage-standards.org.uk/midas/docs/
[155] http://edina.ac.uk/projects/geoxwalk/geoparser.html

projects, that scholars aren't doing *qualitatively* new work but are simply answering old questions more efficiently with new tools.

The CSHE report concluded that in order to support more archaeological scholars interested in doing digital scholarship both training and technical support would be required, but it would also need to reflect the varying capabilities of scholars:

> Many look to their institutions to provide them with support and resources for digital scholarship, but are unable to pay for the services of local technical staff. Digital humanities facilities at some institutions support innovative scholars, but these institutions may be too advanced for the needs of many of the scholars we interviewed and, consequently, have limited uptake by faculty. Some scholars, however, observed that it is easier to get technical help from their institutions if the projects might produce transferable tools and technologies (Harley et al. 2010, pg. 125).

The ability to provide both "simple" and advanced levels of technical assistance is thus required. In addition, building tools that can be repurposed was listed as one way of garnering greater institutional support. One project that has explored the issues of implementing new technology and digital methods for archaeological field research is the Silchester Roman Town [156] project, a British research excavation project of the Roman town of Silchester from its history before the Roman conquest until it was abandoned in the fifth century A.D.

As part of their work, the project made extensive use of a specialized database called the Integrated Archaeological Database (IADB) [157] that was first developed in the 1980s and is now available as a web-based application that makes use of Ajax, MySQL and PHP (Fulford et al. 2010). "Crucial to the interpretation of the archaeological record," Fulford et al. reported, "is the IADB's capacity to build the hierarchical relationships (archaeological matrix) which mirror the stratigraphic sequence and enable the capture of composite, spatial plans of the individual context record to demonstrate the changing character of occupation over time" (Fulford et al. 2010). Archaeological data can be viewed as individual records, 2D matrices or as groups of objects.

One major challenge faced during field research is site recording and Fulford et al. observed that the double handling of data was particularly problematic. To deal with this problem, the Silchester Roman project first collaborated with the OGHAM (On-Line Group Historical and Archaeological Matrix) project that was funded by the Joint Information Systems Committee (JISC) [158] and introduced the use of PDAs and rugged tablet computers for field recording. The most significant insight of this first project was that direct network access was "invaluable" particularly in terms of communication and data management. JISC then continued funding this work through the VERA: Virtual Environment for Research in Archaeology [159] project and the initial collaboration was extended to include information and computer scientists. As Baker et al. (2008) described, the VERA project sought "to investigate how archaeologists use Information Technology (IT) in the context of a field excavation, and also for post-excavation Analysis." The project also introduced new tools and technology to assist in "the archaeological processes of recording, manipulating and analysing data."

Baker et al. also underscored that one of the most important parts of the archaeological process is recording "contexts," which have been defined as the "smallest identifiable unit into which the archaeological record can be divided and are usually the result of a physical action" (Baker et al. 2008). As contexts are identified they are given a unique number in a site register and typically the information is recorded on a paper "context card" that will track everything from sketches to data. Context cards are typically filed in an area folder and then eventually entered manually into a database. This process, however, is not without its problems:

> The recorded contexts provide the material to populate the research environment, they are stored in the Integrated Archaeological Data Base (IADB), which is an online database system for managing recording, analysis, archiving and online publication of archaeological finds, contexts and plans. In the past the entry of data on to the IADB has been undertaken manually. There are around 1000 contexts recorded each season, which means that manual input of the data and information is very time consuming (Baker et al. 2008)

One of the challenges of the VERA project therefore was to find a way to both make the process of recording contexts and entering them into the database more efficient. As their ideal, they cite Gary Lock's goal of

---

[156] http://www.silchester.rdg.ac.uk/
[157] http://www.iadb.org.uk/
[158] http://www.jisc.ac.uk/
[159] http://vera.rdg.ac.uk/

archaeological computing, or where "the information flows seamlessly from excavation, through post-excavation to publication and archive" (Lock 2003, as cited in Baker et al. 2008).

The VERA team asked archaeologists to complete diaries while in the field, conducted one to one interviews and a workshop, and implemented user testing with the IADB. In the diary study of 2007, they asked archaeologists about their experience using digital technology during the excavation process. They met with a fair amount of resistance both to keeping diaries and the use of new technology, and many participants noted the unreliability of wi-fi in the field. Another important insight from the interviews was that data quality was of the highest importance and some archaeologists observed that the direct entry of contexts into the database was often leading to lower quality data. In the first excavation, field workers were entering data directly into the IADB without the traditional quality control layer, but in the second field season, paper recording of contexts was reintroduced with the "small finds supervisor" collating context reports and entering them herself. The diary study thus "illustrated the importance of maintaining existing mechanisms for checking and controlling data" (Baker et al. 2008). Nonetheless, the use of digital pens was rated very highly and did speed entering of some contexts into the IADB. [160] At the same time, the pens were not able to digitally capture all data (43%) from the excavation season directly into the IADB on the first pass (Fulford et al. 2010, pg. 25). This study thus also demonstrated the importance of a willingness to both try and possibly abandon new technologies and to test real disciplinary workflows in the development of the VRE.

Interviews and user testing of the IADB also provided the VERA team with other important information, including the need to make the interface more intuitive and for the database design team to warn users before implementing broad system changes. The three themes that Baker et al. reported ran through all of their research were the need for data quality, transparency of data trails, and "ease of use of technologies."

Fulford et al. (2010) concluded that both VERA and OGHAM had enhanced the work of the Silchester Roman town project:

> The OGHAM and VERA projects have unquestionably strengthened and improved the flow of data, both field and finds records, from the trench to the database, where they can be immediately accessed by the research team. The greater the speed by which these data have become available, the faster the research manipulation of those data can be undertaken, and the faster the consequent presentation of the interpreted field record to the wider research team. The challenge is now to determine whether the same speed can be achieved with the research team of specialist analysts (Fulford et al. 2010, pg. 26).

Although the multidisciplinary project team could both get to their data faster and manipulate their data in new ways, Fulford et al. also reported that it remained to be seen if specialists would begin to publish their results any faster, or if they would publish them electronically. Nonetheless, remote access to the IADB was found to be especially important for specialists, as it allowed them to "become more integrated with the context of their material" and enabled both new levels of independent and collaborative work between specialties.

Additionally, while Fulford et al. stated that the IADB formed the "heart of the VRE" they also envisioned a VRE for archaeology that could support both the digital capture and manipulation of finds from the field and support more sophisticated levels of post-excavation analysis. The Silchester project hoped to publish for both a specialist and public audience, and while acknowledging the new opportunities of electronic publication also stated that they were simultaneously publishing their results in print as well. "Despite the potential of web-based publication, lack of confidence in the medium- or the longer-term sustainability of the web-based resource has meant a continued and significant reliance on traditional printed media," Fulford et al. explained, "This is as much true of Silchester as it is of archaeology in general" (Fulford et al. 2010, pg. 28). At the same time, the Silchester Roman town project has created a constantly evolving project website, a printed monograph, and also published an article in *Internet Archaeology* [161] that referenced data they had archived in the ADS. [162] The creation of a website has also brought with it the concurrent challenges of accessibility and sustainability. Nonetheless, this project demonstrated how digital publishing offered new opportunities while not necessarily precluding print publishing as well. The archiving of their data with the ADS and the publishing of their article

[160] For greater detail on these studies, please see (Warwick et al. 2009).
[161] http://intarch.ac.uk/journal/issue21/silchester_index.html
[162] http://ads.ahds.ac.uk/catalogue/archive/silchester_ahrc_2007/

in *Internet Archaeology* that linked to this data occurred under the auspices of the LEAP (Linking Electronic Archives and Publications) project and demonstrated not only the importance of the long-term preservation of data but the potential of electronic publication for linking archaeological research to the actual data on which it is based.

While the IADB presented a number of opportunities, it also raised three major technical challenges: security on the web, interoperability with other databases, and the potential of 3D visualizations or reconstructions for both academic and public users of the data. On this third point, Fulford et al. asserted that "integral to this is the need to link the evidence used to build the reconstruction with data stored in the IADB." In addition, the need to illustrate to users that all visualizations of Silchester on the website are based on human interpretation of available data was also cited as essential. The reality of human interpretation needs to be carefully considered in any VRE design for archaeology according to Stuart Dunn:

> In the discussion of VREs, and models of data curation and distribution which are based on central and/or institutional storage and dissemination of data, it is easy to forget the interpretive implications of handling archaeological information digitally….this is [SIC] also pertains to the broader arts and humanities VRE agenda. The act of publishing a database of archaeological information implicitly disguises the fact that creating the database in the first place is an interpretive process (Dunn 2009).

Dunn's warning is an important reminder that the design of any infrastructure in the humanities must take into account the interpretative nature of most humanities scholarship. He also further detailed how archaeological workflows are "idiosyncratic, partly informal, and extremely difficult to define," all factors that make them hard to translate into a digital infrastructure (Dunn 2009).

To further this point some recent research using topic modeling and an archaeological database has recently illustrated just how subjective the human interpretations of archaeological data can be. Recent work by David Mimno (2009) used topic modeling and a database of objects discovered in houses from Pompeii [163] to examine the validity of the typological classifications that were initially assigned to these objects. This database contains 6000 artifact records for finds in 30 architecturally similar houses in Pompeii, and each artifact is labeled with one of 240 typological categories and the room in which it was found. Due to the large amount of data available, Mimno argued that the use of statistical data mining tools could help provide some new insights into this data:

> In this paper we apply one such tool, statistical topic modeling, ... in which rooms are modeled as having mixtures of functions, and functions are modeled as distributions over a "vocabulary" of object types. The purpose of this study is not to show that topic modeling is the best tool for archeological investigation, but that it is an appropriate tool that can provide a complement to human analysis. To this aim, we attempt to provide a perspective on several issues raised by Allison, that is, if not unbiased, then at least mathematically concrete in its biases (Mimno 2009).

In common archaeological practice, Mimno explained, artifacts that are excavated are typically removed to secure storage and while their location is carefully noted in modern digs, artifacts in storage are typically analyzed "in comparison to typologically similar objects rather than within their original context." Consequently, Mimno reasoned that determining the function of many artifacts had been driven by "arbitrary tradition" and the perception of individual researchers in terms of what an artifact resembles. Two classes of artifacts, in fact, the "casseruola" (casserole dish) and "forma di pasticceria" (pastry mold) were named based on similarities to 19th century household objects and the creator of the Pompeii database (Penelope Allison) contended that many modern archaeologists made often un-validated assumptions about objects based on their modern names.

For these reasons, Mimno decided to use topic modeling to reduce this bias and explored the function of artifact types using only object co-occurrence data and no typology information. All object descriptions were reduced to integers and then a statistical topic model was used to detect "clusters of object cooccurrence" that might indicate functions. While Mimno admitted that this system still relied on experts having accurately classified physical objects into appropriate categories in the first place, no other archaeological assumptions were made by the training model. The basic assumption was that if two objects shared similar patterns of use they should have

[163] http://www.stoa.org/projects/ph/home

a high probability of cooccurence together in one or more "topics." Initial analysis of a topic model for the "casseruola" and "forma di pasticceria" illustrated them as having little connection to other food preparation objects and thus supported Allison's claim that the modern names for these items are incorrect. This work illustrates how computer science can make it possible for scholars to re-analyze large amounts of existing legacy archaeological data to make new arguments about that data.

Visualization & 3D Reconstructions of Archaeological Sites

The nature of 3D modeling and digital reconstruction has made this particular area of archaeological research one of the most collaborative and groundbreaking. As Harley et al. (2010) have noted, these new approaches have both their benefits and their challenges:

> 3D modeling—based on the laser scanning of archaeological sites, the photogrammetric analysis of excavation photographs, or other virtual modeling techniques—provides unique opportunities to virtually represent archaeological sites. These virtual models do not yet provide a facile publishing platform, but they may allow scholars to run experiments or test claims made in the scholarly literature. Although 3D modeling is a new dimension for archaeological research and dissemination, it may not be suitable for all applications (such as poorly preserved archaeological sites). In addition, some scholars observed that the focus on 3D modeling as a new technology may come at the expense of close attention to physical and cultural research (Harley et al. pg 115).

As with other new technologies, scholars were often worried that a focus on technology would replace exploration of more traditional archaeological questions. Nonetheless, the number of archaeological websites making use of 3d modeling and virtual reconstruction is growing continuously [164] and this section will briefly look at several of the larger projects and the growing body of research literature in this area.

Alyson Gill recently provided a brief overview (Gill 2009) of the use of digital modeling in both archaeology and humanities applications. She reported that Paul Reilly first coined the term "virtual archaeology" in 1991 and since that time his initial concept of three-dimensional modeling of ancient sites has expanded greatly. Instead of virtual archaeology, Koller et al. (2009) suggest a more expansive definition of "virtual heritage" as a:

> …relatively new branch of knowledge that utilizes information technology to capture or represent the data studied by archaeologists and historians of art and architecture. These data include three-dimensional objects such as pottery, furniture, works of art, buildings, and even entire villages, cities, and cultural landscapes (Koller et al. 2009).

The first major book published on this subject, according to both Gill (2009) and Koller et al. (2009), was *Virtual Archaeology*, which was published in 1997 by M. Forte and A. Siliotti. Koller et al. indicated that this book illustrated how early models typically served illustration purposes only and that early publications focused on methodologies used to create such models. In addition, commercial companies created almost all of the models in this book. Since that time, Koller et al. reported, things have changed greatly, the price of 3D modeling software and data capture technology has dropped drastically and the skill sets required to work with these tools is growing among scholars. Gill (2009) has identified four major trends in current projects: collaborative virtual environments, online applications used for teaching and learning, reconstruction of large scale historical spaces, and digital preservation of cultural heritage sites.

One of the largest and perhaps best-known sites described by Gill is the Digital Karnak Project, [165] an extensive website created under the direction of two scholars at the University of California Los Angeles. The temple of Karnak in Egypt existed for over 3000 years and this website has created a number of ways for users to explore its history. A 3-D virtual reality model of the temple was created that allows users to view how the temple was constructed and modified throughout time and this is accompanied by original videos, maps and thematic essays written by Egyptologists. A simplified version of the model of the temple was also made available on Google Earth. There are four ways to enter the website: 1) a Timemap of the site that allows users to choose a time period and view features that were created, modified and destroyed; 2) choosing one of a series of thematic topics with essays and videos; 3) browsing the archive by chronology, type, feature or topic, which takes the

[164] Some interesting sites not covered here include the Skenographia Project (http://www.kvl.cch.kcl.ac.uk/wall_paintings/introduction/default.htm), the Digital Pompeii Project (http://pompeii.uark.edu/Digital_Pompeii/Welcome.html), the Portus Project (http://www.portusproject.org/aims/index.html), and Parthenon 360 (1.) QTVR (http://www.dkv.columbia.edu/vmc/acropolis/#1_1)

[165] http://dlib.etc.ucla.edu/projects/Karnak/

user to both reconstruction model renderings, descriptions in the object catalog, videos, and a large number of photographs; 4) using Google Earth to view the model. This extensive website demonstrates how many of these technologies are being put to use to create sophisticated teaching resources.

The city of Rome has also been the subject of a number of virtual reconstruction projects, with the Digital Roman Forum exploring one particular monument, [166] while the Plan de Rome, [167] the Stanford Digital Forma Urbis Romae [168] project and the particularly well-known Rome Reborn [169] all focus on the city as a whole. To begin with, the Digital Roman Forum provides access to a digital model of the Roman Forum as it appeared in late antiquity and was created by the UCLA Cultural Virtual Reality Lab (CVRLab). [170] Users can use TimeMap to view different features (e.g. the Basilica Aemilia or the Curia Iulia) on the model and clicking on a feature brings up both a virtual model and current photograph of that feature, each of which can have its point of view adjusted. The digital reconstructions can also be searched by keyword or browsed by the primary sources that described it, as well as by function or type. One facet of this website that is particularly noteworthy is that it seeks to integrate the textual sources (such as the histories of Livy and Tacitus) and secondary scholarly research that were used in making some modeling decisions. Each feature also includes a full description [171] with an introduction, history, reconstruction issues including sources and levels of certainty, a bibliography, a series of QuickTime object and panorama movies, and still images. This website illustrates the complicated nature of creating reconstructions, including the amount of work involved, the number of sources used, and the uncertain nature of many visualization decisions.

The Stanford Digital Forma Urbis Romae project provides digital access to the remains "Forma Urbis Romae" a large marble plan of the city that was carved in the third century A.D. This website includes digital photographs, 3D models of the plan, and a database that includes details on all of the fragments. Similarly, the Plan de Rome website provides access to a virtual 3D model of the "Plan of Rome," a large plaster model of the city that was created by architect Paul Bigot, and provides an extraordinary level of detail on the city.

The most ambitious of all of these projects, Rome Reborn, is an international effort that seeks to create 3D models that illustrates the urban development of Rome from the late Bronze age (1000 B.C.) to the early Middle Ages. They have decided to focus initially on 320 A.D. because at this time Rome had reached its peak population, many major churches were being built, and few new buildings were created after this time. A number of partners are involved in the effort including the Institute for Advanced Technology in the Humanities (IATH) of the University of Virginia and the CVRLab. Among the many goals of the project, the website notes:

> The primary purpose of this phase of the project was to spatialize and present information and theories about how the city looked at this moment in time, which was more or less the height of its development as the capital of the Roman Empire. A secondary, but important, goal was to create the cyberinfrastructure whereby the model could be updated, corrected, and augmented.

Currently, a large number of reconstruction stills can be viewed at the website, and in November 2008 a version of Rome Reborn 1.0 was published on the Internet through Google Earth. [172]

Guidi et al. (2006) have reported on one of the most significant efforts in creating Rome Reborn, the digitizing of the Plastico di Roma Antica, a physical model of Rome that is owned by the Museum of Roman Civilization and was designed by Italo Gismondi. The Plastico is a huge physical model of imperial Rome with a high level of intricate detail and creating a digital model of it required the development of a number of advanced imaging techniques and algorithms. Ultimately, the digitized Plastico was used as the basis for a hybrid model of late antique Rome that was also based on new born-digital models created for specific building and monuments in the city. The sheer size of their project required utilizing such a model, for as the authors noted:

---

[166] http://dlib.etc.ucla.edu/projects/Forum
[167] http://www.unicaen.fr/services/cireve/rome/index.php?langue=en
[168] http://formaurbis.stanford.edu/docs/FURdb.html
[169] http://www.romereborn.virginia.edu/
[170] http://www.cvrlab.org/
[171] http://dlib.etc.ucla.edu/projects/Forum/reconstructions/CuriaIulia_1
[172] http://earth.google.com/rome/

> Modeling of an ancient building may start from the historical documentation, archeological studies undertaken in the past and sometimes from a new survey of the area. These data are then combined in the creation of a digital three-dimensional (3D) synthesis that represents a reasonable hypothesis of how the artifact once appeared. The construction of an entire city can proceed by repeating this method as long as needed, but the process would of course be extremely time-consuming, assuming it would be at all possible since sometime (as in the case discussed in this paper) all the archaeological data that would be needed are not known (Guidi et al. 2006).

One interesting idea also suggested by Guidi et al. was that the Plastico is but one physical model of a city and that there have been hundreds of such models developed since the Renaissance, so the methodologies they have used could easily be transferred to the development of digital models of other cities.

While smaller in scale than Rome Reborn, one of the longest running virtual reconstruction projects is the Pompey Project, [173] which has developed an extensive website that includes a history of the theatre of Pompey, an overview of classical theatre, details on historic and modern excavations at this site with extensive images, and a series of 3d visualizations of the Pompey theatre. Beacham and Denard (2003) provide both a practical and theoretical overview on creating digital reconstructions of the theatre of Pompey, and also examine some of the issues such reconstructions create for historical study. One of the greatest advantages of virtual modeling technology they found was its ability to integrate "architectural, archaeological, pictorial and textual evidence" to create new 3-dimensional "virtual performance spaces." [174]

The use of 3d modeling, Beacham and Denard observed, allowed them to manipulate huge datasets of different information types and this in particular supported better hypotheses in terms of "calculating and documenting degrees of probability in architectural reconstructions." Nonetheless, the authors also stressed that the data used in such models must be carefully evaluated and coordinated. At the same time, virtual models can both be updated more quickly than traditional models or drawings with new information as it becomes available and represent alternative hypotheses. The authors argued that the creation of a website thus supports a more sophisticated form of publication that allows for rapid dissemination of scholarly information that can be continuously updated with new information.

The ability to represent multiple hypotheses and to provide different reconstructions the authors also concluded supports the "liberation" of the reader, so they can "interpret and exploit the comprehensive data according to their own needs, agendas and contexts." The nature of this work is also inherently interdisciplinary, and involves scholars in multiple disciplines as well as various technicians. Beacham and Denard ultimately argued that digital reconstructions are inherently a new form of scholarship:

> The very fact that this work is driven by the aim of creating a three-dimensional reconstruction of the theatre has, itself, far-reaching implications. The extrapolation of a complete, three-dimensional form from fragmentary evidence, assorted *comparanda* and documentary evidence is quite different in character to the more frequently encountered project of only documenting the existing remains of a structure (Beacham and Denard 2009).

The authors also warned, however, that digital reconstructions must avoid the lure of the "positivist paradigm" or in other words, digital models should never be presented as "reality." All reconstructions must thus be considered as varying hypotheses with different levels of probability and this must be made very clear to the user, or else the utility of these models as teaching and scholarly communication tools is dubious at best.

The utility of reconstructions and models in teaching has been explored extensively by the Ashes2Art project, [175] a collaboration between Coastal Carolina University in South Carolina and Arkansas State University in Jonesboro, where students create 3-dimensional computer models of ancient monuments based on excavation reports, build educational and flythrough videos, take on-site photographs of architectural details, write essays, create lesson plans, and ultimately document all of their work online with primary and secondary source bibliographies (Flaten 2009). The development of the Ashes2Art collaboration provides an innovative example of undergraduate research, faculty-student collaboration, and the development of an online resource for both specialists and the general public.

---

[173] http://www.pompey.cch.kcl.ac.uk/

[174] Other recent work has gone even further and has tried to repopulate ancient theatre reconstructions with human avatars (Ciechomski et al. 2004).

[175] http://www.coastal.edu/ashes2art/projects.html

While the first iteration of the course had students working in all of the different areas, the instructors soon realized this was overly ambitious for a semester long course and students were grouped by areas of interest (developing models, designing or updating the Web platform, essay writing and preparing bibliographies, preparing teaching materials, creating videos). Nonetheless all of these groups depended on each other for the final product. At the end of the semester a panel of external scholars reviewed all the models. Although the development of models was the end goal of the Ashes2Art project, the course also addressed larger issues regarding digital models and the reconstruction of archaeological artifacts. As summarized by Flaten:

> The opportunity to visualize complex dimensional data has never been greater, but digital reconstructions and models are not without their critics. Questions of accuracy, methodology, transparency, accessibility, availability, and objective peer review are legitimate concerns (Flaten 2009).

Similar to Beacham and Denard (2003), Flaten emphasized the importance of publishing levels of certainty regarding the data used to create a reconstruction, making all of the data that was utilized explicit to the user, and informing the user that multiple interpretations are possible. Flaten commented that his students were creating "perception models" rather than "structural models." Students used a variety of data including published excavation reports, journal articles, and photography to create general structural models. When there were conflicting accounts in the data and decisions had to be made, both the evidence and the decision made were recorded along with other metadata for the model. Flaten also reiterated that the ability to update digital reconstructions as new information becomes available is one of the greatest strengths of the digital approach. Another important insight gained from this process Flaten observed was that students "discover that uncertainty is a crucial component of knowledge, that precision does not imply accuracy, and that questions are more important than definite answers." Through their work on Ashes2Art, students learned important lessons about how scholarly arguments are constructed and that the creation of new knowledge is always an ongoing conversation rather than a finished product.

As this section has demonstrated, there are a large number of significant archaeological projects exploring the use of 3d models and digital reconstruction. Nonetheless, the ability to preserve these projects and provide long-term access to them is an issue that has received little attention. Koller et al. (2009) have proposed creating open repositories of authenticated 3d models that are based on the model of traditional scholarly journals. [176] Such repositories must include mechanisms for peer review, preservation, publication, updating, and dissemination. In addition to the lack of a digital archive, the authors also criticized the fact that there is no central finding tool to even discover if a site or monument has been digitally modeled, making it difficult to either repurpose or learn from other scholars work. This state of affairs led them to the following conclusion:

> A long-term objective, then, should be the creation of centralized, open repositories of scientifically authenticated virtual environments of cultural heritage sites. By scientifically authenticated, we mean that such archives should accession only 3D models that are clearly identified with authors with appropriate professional qualifications, and whose underlying design documents and metadata are published along with the model. Uncertainties in the 3D data and hypotheses in the reconstructions must be clearly documented and communicated to users (Koller et al. 2009).

The ability to visualize uncertainty in digital models and present these results to users is also a significant technical challenge the authors reveal, and one that has been the subject of little if any research. The development of such repositories, however, face a number of research challenges including: digital rights management for models, uncertainty visualization in 3D reconstructions, version control for models (e.g. different scholars may generate different versions of the same model, models change over time), effective metadata creation, digital preservation, interoperability, searching across 3D models, the use of computational analysis tools with such a repository, and last but by no means least, the development of organizational structures to support them.

---

[176] The creation of such a repository is part of their larger "SAVE: Serving and Archiving Virtual Environments" project (http://www3.iath.virginia.edu/save/) which when complete "will be the world's first on-line, peer-reviewed journal in which scholars can publish 3D digital models of the world's cultural heritage (CH) sites and monuments." On a larger scale, the non-profit CyArk High Definition Heritage Network (http://archive.cyark.org/) is working "to digitally preserve cultural heritage sites through collecting, archiving and providing open access to data created by laser scanning, digital modeling, and other state-of-the-art technologies."

Classical Art & Architecture

The diverse world of classical art and architecture is well represented online and this section will briefly survey a number of specific digital projects. To begin with, one impressive website entitled the "Ancient Theatre Archive: A Virtual Reality Tour of Greek and Roman Theatre Architecture" [177] has been created by Professor Thomas G. Hines of Whitman College and is an excellent resource that can be used to study ancient Greek and Roman theatres. This website provides both a list and graphical map overview for navigating through images of classical theatres. Each theatre page includes an extensive history, a timeline and a virtual tour that includes panorama images and Quicktime movies. A recent addition is a table of "Greek and Roman Theatre Specification" that includes extensive details on each theatre including its ancient name, modern name, location, date, width, capacity, renovation dates, and summary. While this table can be sorted by type of data, it would also have been useful to hyperlink the theatres to their descriptive pages. This website is an excellent educational resource, but it does not seem, unfortunately, that any of the extensive historical data compiled, image or video data can be either downloaded in any kind of standard format or reused.

Many websites are dedicated to the architecture of individual buildings or cities. For example, Trajan's Column, [178] hosted by McMaster University, is dedicated to the exploration of the column of Trajan as a sculptural monument. This website includes introductory essays, a database of images and useful indices for the website. Although over 10 years old this website still stands up well as an educational resource, and even more importantly perhaps, provides technical details on its creation and all of the source code used in its creation. [179]

Another useful tool for browsing across a number of classical art objects is the Perseus Art & Archaeology Browser [180] that allows the user to browse the digital library's image collection including coins, vases, sculptures, sites, gems and buildings. The descriptions and images have been produced in collaboration with a large number of museums, institutions and scholars. In one interesting related project, 3D models were developed from the photographs of vases in this collection and were used to build a "3D Vase Museum" that users could browse (Shiaw et al. 2004). Although the entire collection of art objects can be searched, the major form of access to the Art & Archaeology collection is provided through a browsing interface, where the user must pick an artifact type such as a coin, a property of that artifact (such as material), a property of that artifact type (such as bronze) and this then leads to a list of images. [181] Each catalog entry includes descriptive information, photographer credits and the source of the object and photograph. In addition, each image in the Art & Archaeology Browser has a stable URL for linking. All of the source code used to create this browsing environment is available for download as part of the Perseus Hopper. [182]

The largest research effort in making classical art available online is CLAROS (Classical Art Research Online Services), [183] a major international interdisciplinary research initiative that plans to release a significant online classical art resource by the end of 2010. While further details on this project are discussed later in this paper, this section will consider some of the larger research questions addressed by CLAROS and described in (Kurtz et al. 2009). This project is led by Oxford University and hosted by the Oxford e-Research Center and its data web will integrate the collections of Arachne (Research Archive for Ancient Sculpture-Cologne), [184] the Beazley Research Archive, [185] the Lexicon Iconographicum Mythologicae Classicae (LIMC Basel [186] and LIMC Paris [187]), the German Archaeological Institute (DAI), [188] and the LGPN (Lexicon of Greek Personal Names). [189]

---

[177] http://www.whitman.edu/theatre/theatretour/home.htm
[178] http://cheiron.mcmaster.ca/~trajan/
[179] http://cheiron.mcmaster.ca/~trajan/tech.html
[180] http://www.perseus.tufts.edu/hopper/artifactBrowser
[181] http://www.perseus.tufts.edu/hopper/artifactBrowser?object=Coin&field=Material&value=Bronze
[182] http://www.perseus.tufts.edu/hopper/opensource/download
[183] http:// www.clarosweb.org
[184] http://www.arachne.uni-koeln.de/
[185] http://www.beazley.ox.ac.uk/
[186] http://www.limcnet.org/Home/tabid/77/Default.aspx
[187] http://www.mae.u-paris10.fr/limc-france/
[188] http://www.dainst.org/
[189] http://www.lgpn.ox.ac.uk/

All these founding members have extensive datasets on antiquity in varying formats with more than 2,000,000 records collectively.

CLAROS plans to use the CIDOC-CRM [190] ontology to integrate these different collections and is including the LGPN to place classical "art in its ancient cultural context" and provide "a natural bridge to the large and well developed epidoc community" (Kurtz et al. 2009). While data integration presents a number of difficulties, the authors emphasize that:

A guiding principle of CLAROS is that no partner should need to change the format of his data to join. Each of the founder members uses different databases and front end programs for entering, querying and displaying results through their own websites. Data have been exported from each partner into a common CIDOC CRM format (Kurtz et al. 2009).

Consequently, this approach presents the challenge of providing a "data web" that supports searching across five different collections while still permitting individual organizations to maintain their own databases with their own unique standards.

The CLAROS data web represents each resource as a SPARQL endpoint that is then queried by the SPARQL RDF-query language and returns data as RDF. [191] The two main problems of this approach, as reported by Kurtz et al. (2009) are semantic integration (alignment of different data schemas or ideally mapping to a single schema or ontology) and co-reference resolution (ensuring a reference to the same object or entity in different databases with different names) is recognized as such. Like EAGLE and LaQuAT for epigraphy and APIS for papyrology, CLAROS seeks to provide a federated database search for multiple classical collections. It is their hope that such an architecture will allow them to integrate additional classical art collections by mapping their unique schemas to the core CIDOC-CRM ontology of CLAROS and adding all necessary entries in the co-reference service.

The CLAROS project has many goals including making digital facsimiles of images and reference works available to the general public, providing scholars access with "datasets of intellectually coherent material easily and swiftly through one multi-lingual search facility," enabling museums to access records about both their and other museum collections, and finally, to permit both the public and educational institutions in particular to engage with "high art" in new ways. One particular new way of engaging that they have developed is a prototype that allow users to query for new images by picking an initial image, such as finding all vases similar to the one they have selected based only on image recognition and not textual descriptors. Another intriguing vision of this project is one where members of the public could take images of classical art around the world and upload them to "CLAROS" clouds for image recognition, identification and documentation by experts.

Classical Geography

The potential of digital technologies for the study of geography, within the larger context of the digital humanities and with a specific focus on classical geography and archaeology, [192] has recently been provided by Stuart Dunn (Dunn 2010). Dunn noted that digital technologies are supporting what he labeled "neogeography" a "discipline" that is collaborative, includes new sets of tools and methods, but also presents its own challenges:

The 'grand challenge' for collaborative digital geography therefore, with its vast user base, and its capacity for generating new data from across the specialist and non-specialist communities, is to establish how its various methods can be used to *understand* better the construction of the spatial artefact, rather than simply to *represent* it (Dunn 2010, pg. 56).

---

[190] CIDOC-CRM is a conceptual reference model that "provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation" (http://www.cidoc-crm.org/) and was designed to both promote and support the use of a "common and extensible semantic framework" by various cultural heritage organizations including museums, libraries and archives. For more on the CIDOC-CRM and its potential for supporting semantic interoperability between various digital resources and systems see (Doerr and Iorizzo 2008).

[191] http://www.w3.org/TR/rdf-sparql-query/

[192] A number of public domain reference works have been digitized and are useful for the study of classical geography, such as the *Topographical Dictionary of Ancient Rome* (http://www.lib.uchicago.edu/cgi-bin/eos/eos_title.pl?callnum=DG16.P72) that has been put online by the University of Chicago and the *Tabula Peutingeriana*, a medieval copy of a Roman map of the empire, http://www.hs-augsburg.de/~harsch/Chronologia/Lspost03/Tabula/tab_intr.html

This challenge of not just digitizing or representing traditional objects of study online but finding new ways to use these methods to conduct innovative research, create new knowledge and answer new questions is a recurrent theme seen through the disciplines of digital classics. Several prominent digital projects that focus on addressing these challenges within the realm of classical geography shall provide the focus of this section.

The Ancient World Mapping Center

The Ancient World Mapping Center (AWMC), [193] an interdisciplinary center at the University of North Carolina-Chapel Hill, is perhaps the preeminent organization in this field of study. According to its website, the AWMC "promotes cartography, [194] historical geography and geographic information science as essential disciplines within the field of ancient studies through innovative and collaborative research, teaching, and community outreach activities." This website includes a number of resources for researchers but the majority of the website is composed of short research articles written by AWMC staff regarding various topics such as new publications available or websites of interest and these can be found by browsing the table of contents for the website, the topical index or by searching. There is also an index of place names used in articles on the website. One other useful resource is a selection of free maps of the classical world that can be downloaded for teaching.

The Pleiades Project

The Pleiades Project, [195] once solely based at the AWMC but now a joint project of the AWMC, the Institute for the Study of the Ancient World (ISAW), [196] and the Stoa Consortium, [197] is one of the largest digital resources in classical geography. The Pleiades website allows scholars, students and enthusiasts to both share and use historical geographical information about the classical world. A major goal of Pleiades is to create an authoritative digital gazetteer of the ancient world that is continuously updated and can be used to support other digital projects and publications through the use of "open, standards based interfaces" (Elliott and Gillies 2009a).

From its very beginning, Pleiades was intended to be collaborative and in order to join the project and contribute information a user simply needs an email address and to accept a contributor agreement. This agreement leaves all intellectual property rights with contributors but also grants to Pleiades a "CC Attribution Share Alike License." Registered users can suggest updates to geographic names, add bibliographic references and contribute to descriptive essays. While all contributions are vetted these suggestions then "become a permanent, author-attributed part of future publications and data services" (Elliott and Gillies 2009b). Thus the Pleiades project provides a light level of "peer-review" to all user contributed data.

The content within the Pleiades gazetteer "combines "pure" data components (e.g., geospatial coordinates) with the products of analysis (e.g., toponymic variants with indicia of completeness, degree of reconstruction and level of scholarly confidence therein) and textual argument" (Elliott and Gillies 2009a). In addition, Pleiades also includes content from the Classical Atlas project, an extensive international collaboration that led to the publication of the *Barrington Atlas of the Greek and Roman World.* In fact, the creators of Pleiades see the website as a permanent way to update information in the Barrington online. The creators of Pleiades consider their publication model as in some ways close to both an academic journal and encyclopedia:

> Instead of a thematic organization and primary subdivision into individually authored articles, Pleiades pushes discrete authoring and editing down to the fine level of structured reports on individual places and names, their relationships with each other and the scholarly rationale behind their content. In a real sense then Pleiades is also like an encyclopedic reference work, but with the built-in assumption of on-going revision and iterative publishing of versions (Elliott and Gillies 2009b).

---

[193] http://www.unc.edu/awmc/

[194] One excellent interactive resource for studying the cartographic history "of the relationships between hydrological and hydraulic systems and their impact on the urban development of Rome, Italy" is *Aquae Urbis Romae* (http://www3.iath.virginia.edu/waters/)

[195] http://pleiades.stoa.org/

[196] ISAW (http://www.nyu.edu/isaw/) is based at New York University (NYU) and is a "center for advanced scholarly research and graduate education, intended to cultivate comparative and connective investigations of the ancient world from the western Mediterranean to China, open to the integration of every category of evidence and relevant method of analysis" and will feature a variety of doctoral and postdoctoral programs to support groundbreaking and interdisciplinary scholarship.

[197] http://www.stoa.org/

Rather than using toponyms or coordinates as the primary organizing theme of the website, they have used the concept of place "as a bundle of associations between attested names and measured (or estimated) locations (including areas)" (Elliott and Gillies 2009b). These bundles are then called features, which can be positioned in time and have scholarly confidences registered to them. The ability to indicate levels of confidence in historical or uncertain data is an important part of many digital classics projects.

As the sheer amount of content is far beyond the scale of individual project participants to actively edit and maintain, the Pleiades project has "pushed out" this responsibility to interested members of the classics community and beyond, through the collaboration model described above. Another important feature of Pleiades Pleiades is that only uses open source software such as OpenLayers, [198] Plone, [199] and zgeo. [200]

In addition, the Pleiades project promotes the use of their gazetteer as an "authority list" for Greek and Roman geographic names and their associated locations. All Pleiades content has stable URLs for its discrete elements and this allows other digital resources to "refer unambiguously to the places and spaces mentioned in ancient texts, the subjects of modern scholarly works, the minting locations of coins, and the findspots of inscriptions, papyri, and the like" (Elliott and Gillies 2009a). The difficulties that historical place names within "legacy literature" and the challenges they present to named entity disambiguation, geoparsing and automatic mapping techniques was previously reported by the Archaeotools project, and as Elliott and Gillies (2009b) describe is also a significant challenge for classical geography. [201] They detailed how many historical books found within Google Books have information pages that include Google Maps populated with place names extracted from the text, but classical place names such as Ithaca, however, are often assigned to modern places such as the city in New York by mistake. While there are algorithms that attempt to deal with many of these issues, they also argue that:

> This circumstance highlights a class of research and publication work of critical importance for humanists and geographers over the next decade: the creation of open, structured, web-facing geo-historical reference works that can be used for a variety of purposes, including the training of geo-parsing tools and the population of geographic indexes (Elliott and Gillies 2009b).

Part of the research of the Pleiades project, therefore, has been to determine how best to turn digital resources such as their gazetteer into repurposeable knowledge bases. Elliott and Gillies (2009b) predict that increasingly those who hold geographic data and wish to make it freely available online will provide access to their data through a variety of web services.

Despite their desire to make all of Pleiades content available to be remixed and mashed up, these efforts have met with some obstacles:

> In our web services, we employ proxies for our content (KML [202] and GeoRSS [203]-enhanced Atom [204] feeds) so that users can visualize and exploit it in a variety of automated ways. In this way, we provide a computationally actionable bridge between a nuanced, scholarly publication and the geographic discovery and exploitation tools now emerging on the web. But for us, these formats are lossy: they cannot represent our data model in a structured way that preserves all nuance and detail and permits ready parsing and exploitation by software agents. Indeed, we have been unable to identify a standard XML-based data format that simply and losslessly supports the full expression of the Pleiades data model (Elliott and Gillies 2009b).

In order to provide a lossless export, they plan to produce file sets composed of ESRI shape files with attribute tables in CSV, a solution that despite the proprietary nature of the ShapeFile format, does allow them to download time-stamped files into the institutional repository at New York University. Although they would

---

[198] OpenLayers is an open source JavaScript library that can be used to display map data in most Web browsers (http://openlayers.org/)
[199] Plone is an open source content management system (http://plone.org/).
[200] http://plone.org/products/zgeo.wfs
[201] The use of computational methods and customized knowledge sources for historical named entity disambiguation has an extensive literature that is beyond the scope of this paper, but for some useful approaches see (Smith 2002, Tobin et al. 2008)
[202] KML formerly known as "keyhole markup language" was created by Google and is maintained as an open standard by the Open Geospatial Consortium(http://www.opengeospatial.org/standards/kml/). KML is an "XML language focused on geographic visualization, including annotation of maps and images" and is utilized by a number of open-source mapping projects. (http://code.google.com/apis/kml/documentation/mapsSupport.html)
[203] GeoRSS (http://www.georss.org/Main_Page) is a "lightweight, community driven way to extend existing feeds with geographic information"
[204] According to Wikipedia (http://en.wikipedia.org/wiki/Atom_(standard)), the name Atom applies to two related standards, while the Atom Syndication Format is a "XML language used for web feed" (http://www.ietf.org/rfc/rfc4287.txt) the Atom Publishing Protocol is a "HTTP-based protocol for creating and updating web resources."

prefer to use only open formats, Elliott and Gillies also argued that the ShapeFile format is used around the world and can be decoded by open-source software, a fact that gives it a "high likelihood of translation into new formats in the context of long-term preservation repositories." The experience of Pleaides illustrates the challenges of wanting to create open access resources while also having to deal with the limits of open formats and long-term preservation needs.

Nonetheless, the open access nature of Pleiades and the ability to link to individual places within it makes it a natural source to integrate with other digital classics projects in numismatics, epigraphy and papyrology, or any digital resource that makes extensive use of historical place names within the ancient world. Indeed, Pleiades is actively working with other projects through the Concordia initiative to integrate its content with other digital projects and "develop standards-based mechanisms for cross-project geographic search."

The HESTIA Project

While smaller in scale than Pleiades, the HESTIA (Herodotus Encoded Space-Text-Imaging-Archive) [205] project provides an interesting look at how digital technology and spatial analysis [206] can be used to answer a specific research question in classical geography. HESTIA seeks to examine the different ways in which the history of Herodotus refers to space and time. [207] Several of their major research questions include studying his "representation of space in its cultural context," exploring if different peoples represented in his history conceive of space differently, and testing the thesis "that the ancient Greek world centered on the Mediterranean and was comprised of a series of networks" (Barker 2010).

Barker et al. (2010) have provided an extensive overview of the design and initial findings of HESTIA, including methodological considerations for other projects that seek to make use of the state-of-the-art in GIS, relational databases and other computational tools to explore questions not just in classical geography but also in the humanities in general. The authors also demonstrate how many traditional questions regarding the text of Herodotus (e.g. what is the relative importance of bodies of water (particularly rivers) in organizing the physical and cultural space?) can be investigated in new ways using digital technologies. The authors also identified a number of themes that HESTIA would pursue in-depth regarding the thinking about space in Herodotus *Histories:*

> …namely, the types of networks present and their interpretation, the influence of human agency and focalisation, the idea of space as something experienced and lived in, and the role of the medium in the representation of space—and to emphasise that close textual reading underpins our use of ICT throughout (Barker et al. 2010).

One major point reiterated continuously by Barker et al. was that a close *textual* reading of Herodotus was the first consideration before making any technological decisions.

The methodology in creating HESTIA involved four stages: 1) utilizing the digital markup of a Herodotus text obtained from Perseus; 2) compiling a spatial database from this text; 3) producing basic GIS, GoogleEarth and Timeline maps with this database and 4) creating and analyzing automated network maps. One particularly interesting feature of the HESTIA project is that it repurposed the TEI-XML versions of Herodotus available from the Perseus Digital Library, and in particular the place names tagged along with coordinates and identifiers from the Getty Thesaurus of Geographic Names (TGN) [208] in the English file. Nonetheless, this process of reuse was not seamless and HESTIA needed to perform some procedural conversions. Specifically, they converted

---

[205] http://www.open.ac.uk/Arts/hestia/index.html

[206] The HESTIA project hosted a workshop in July 2010 entitled "New worlds out of old texts: interrogating new techniques for the spatial analysis of ancient narratives" that will bring together numerous projects that are using spatial analysis techniques in various classical disciplines (http://www.arts-humanities.net/event/new_worlds_out_old_texts_interrogating_new_techniques_spatial_analysis_ancient_narratives)

[207] Explorations of how time and space were conceived of in the ancient world is also the focus of the German research project TOPOI, "The Formation and Transformation of Space and Knowledge in Ancient Civilizations." http://www.topoi.org/index.php. According to their website, this "interdisciplinary research association investigates ancient civilizations from the 6th millennium BC to Late Antiquity. Issues in focus are: How have spatial orders and knowledge developed? How are space and knowledge related?" Recent details on one TOPOI project can be found in (Pappelau and Belton 2009).

[208] http://www.getty.edu/research/conducting_research/vocabularies/tgn/

the TEI P4 file from Perseus to TEI P5 and the Greek text was transformed from Beta code to Unicode using a Transcoder tool developed by Hugh Cayless.

The HESTIA project also decided to use only the English version of the *Histories* to probe spatial data since the Greek text would have to have had toponyms tagged by hand. Nonetheless, since they still wanted to make use of the Greek text, they assigned unique identifiers to each section of the text in Greek and English so that associations could still be made. In addition, the project needed to perform some data cleaning of the geographic markup in the Perseus TEI XML file including removing duplicate entries and correcting coordinates, entity categorizations and references to false locations. This work by HESTIA illustrates that even while the creation of open access texts by digital classics projects supports reuse, this reuse is not without its own computational challenges and costs.

After correcting and standardizing the place name tagging in the English *Histories* of Herodotus, the HESTIA project extracted this information and compiled a spatial database stored in PostgreSQL [209] (which has a PostGIS extension). This database can be queried to produce different automated results that can then be visualized through maps. The generation of this database posed a number of problems, however, including questions as to what if any connections Herodotus might have been drawing between places, the quality of the English translation, and various syntactic issues of language representation. Nonetheless, the final database has a simple structure and contains only three tables: sections (which stores information about the section of Herodotus text), locations (which stores unique locations), and references (this table ties the locations and sections together by providing unique IDs for all references to spatial locations within Herodotus). Whereas Perseus had used only a single level of categorization, HESTIA introduced a broader level of categorization of "geotype" and "subtype," a process that also defines places as settlements, territories or physical features.

The HESTIA project chose QGIS, [210] an open source GIS tool (that connects easily to PostGIS), as the application for querying the database and generating maps. As with the choice of PostgreSQL, Barker et al. were concerned with choosing applications that would support long-term data preservation and analysis. Using these tools allowed SQL queries to be generated that could perform various functions with related maps including producing a gazetteer of sites, listing the total number of references to a location (such as by book of Herodotus), and generating a network based on co-reference of locations (e.g. within a specific book). In order to provide greater public access to this data, the HESTIA project decided to expose the PostGIS data as KML so that it could be read by various mapping applications such as GoogleEarth. The project accomplished this by using GeoServer, an "Open Source server that serves spatial data in a variety of web-friendly formats simultaneously" including KML and SVG. The creation of this "Herodotus geodata" allows users to construct their own mashups of the visual and textual data created by the HESTIA project.

In order to more successfully visualize spatial changes in the narrative, the project made use of TimeMap.js, [211] an open source project that uses several technologies to "allow data plotted on GoogleMaps to appear and disappear as a timeline is moved." The project thus hired the developer of TimeMap, Nick Rabinowitz, to create a "Herodotus Narrative Timeline", that "allows users to visualise locations referred to in the text by scrolling along a 'timeline' representing each chapter." [212] The development of this timeline, however, also required the creation of one feature to better integrate textual and visual data:

> When places are first mentioned, they appear flush to the right-hand side of the 'timeline' bar and fully coloured in on the map. As one moves through the narrative, however, they move to the left of the 'timeline' bar accordingly and become ever fainter on the map, until, in both cases, they drop out altogether. In doing this, we have tried to reproduce more accurately the reading experience: in the mind; some chapters later, this place might no longer hold the attention so greatly, but its memory lingers on (captured in the Timeline Map by the faded icons), until it disappears altogether. By re-visualising the data in this format, we hope not only to assist in the reading experience of Herodotus but also to raise new research questions that would not have been apparent before the advent of such technology (Barker et al. 2010)

[209] http://www.postgresql.org/about/.
[210] http://www.qgis.org/
[211] http://code.google.com/p/timemap/
[212] http://www.nickrabinowitz.com/projects/timemap/herodotus/basic.html

The development of this timeline and reading tools to support its use demonstrates how digital technologies offer new ways of reading a text.

HESTIA has also produced a number of automatic network maps to analyze how the narrative of Herodotus organized space and relations between places. Barker et al. cautioned, however, that as accurate as such maps may appear in GoogleEarth, they can never truly be objective, for "the form chosen to represent space inevitably affects its understanding and is in itself framed by certain epistemological assumptions" (Barker et al. 2010). This warning cogently echoes the fears articulated earlier by scholars who created digital reconstructions of archaeological monuments (Beacham and Denard 2003), that users of such visualizations and maps must always be cognizant of the theoretical and interpretative arguments inherent in such constructions. Nonetheless, the HESTIA project created network maps as a way to better analyze connections made *between* places in the narrative of Herodotus, and they focused on *topological* (links created by human agents) rather than *topographic* connections (two dimensional maps), or as they clarify: "in other words, we are interested in capturing and evaluating the mental image (or, better, images) of the world contained within Herodotus' narrative, not any supposed objective representation"(Barker et al. 2010).

The creators of HESTIA also explained that the creation of automated network maps was in large part designed to increase the impact of the project, or to bring Herodotus to new readers on the web. At the same time, the queries and the maps they generated are intended to prompt *new* questions and analysis rather than provide definitive answers, or as Barker et al. (2010) accentuated they "should be regarded as *complementing* rather than replacing close textual analysis." In fact, the automated network maps faced a variety of problems with spatial data inherited from the Perseus English text as well as the far greater issue that many place names in the English translation and thus in the database are not found in the Greek text. As Barker et al. (2010) explained, one fundamental difference between Greek and English is that often what was conceptualized of as a "named geographical concept" in English was represented in Greek as the *people* who lived in that place. For future work, Barker et al. (2010) posited that they would need to further nuance their quantitative approach with qualitative approaches that mine the textual narrative. In sum, this project illustrates not only how the digital objects of other projects can be repurposed in new ways but also how a traditional research question in classics can be reconceived in a digital environment.

One of the principal researchers of the HESTIA Project has also published other work into the use of technologies such as GIS and network analysis to research questions in classical geography (Isaksen 2008). Leif Isaksen's [213] research into the Roman Baetica integrates the archaeological and documentary record (Antonine Itineraries, Ravenna Cosmography) and demonstrates the potential of using new technologies to revisit arguments about transportation networks and the Roman Empire.

Digital Editions & Text Editing

Introduction

Perhaps one of the most extensive if frequently debated questions in the field of digital classics and indeed in all of the digital humanities is both how to build an infrastructure that supports the creation of "true" digital critical editions, and also what constitutes a "digital critical edition." Several longstanding projects serve as examples of creating digital editions for the entire corpus of an author (Chicago Homer [214]), for the selected works of an author (Homer and the Papyri, [215] Homeric Multitext [216]) or for individual works by individual authors (Electronic Boethius, [217] Ovid's *Metamorphoses*, [218] and the Vergil Project [219]).

[213] Leif Isaksen also recently presented initial work in using digital methods to analyze the *Geographia* of Claudius Ptolemy (Isaksen 2010).
[214] http://www.library.northwestern.edu/homer/
[215] http://chs.harvard.edu/wb/86/wo/XNM7918Nrebkz3KQV0bTHg/4.0.0.0.19.1.7.15.1.1.0.1.2.0.4.1.7.1.0.0.1.3.3.1
[216] http://chs.harvard.edu/wa/pageR?tn=ArticleWrapper&bdc=12&mn=1169
[217] http://beowulf.engl.uky.edu/~kiernan/eBoethius/inlad.htm
[218] http://etext.lib.virginia.edu/latin/ovid/
[219] http://vergil.classics.upenn.edu/

In their discussion of multitexts and digital editions, Blackwell and Crane (2009) offered an overview of digital editions and what an ideal digital library infrastructure might provide for them:

…digital editions are designed from the start to include images of the manuscripts, inscriptions, papyri and other source materials, not only those available when the editor is at work but those which become available even after active work on the edition has ceased…. This is possible because a true digital edition will include a machine actionable set of sigla. Even if we do not yet have an internationally recognized set of electronic identifiers for manuscripts, the print world has often produced unique names (e.g., LIBRARY + NUMBER) that can later be converted into whatever standard identifiers appear. A mature digital library system managing the digital edition will understand the list of witnesses and automatically search for digital exemplars of these witnesses, associating them with the digital edition if and when they come on-line (Blackwell and Crane 2009).

They stated that digital editions need to include images of all of their primary sources of data, and that "true" digital editions will be dynamic and automatically search for digital facsimiles of manuscript witnesses as they become available, provided standard sets of machine actionable identifiers are used.

Whether in a print or digital infrastructure, Borgman has also noted the continuing importance of access to various editions of a text, commenting that humanities scholars and students: "also makes the finest distinctions among editions, printings, and other variants – distinctions that are sometimes overlooked in the transition from print to digital form (Borgman 2009). This section will provide a brief overview of some of the theoretical and technical issues involve in creating a cyberinfrastructure for digital editions in classics and beyond.

Theoretical Issues of Modeling and Markup for Digital Editions

The traditional task of creating a critical edition in classics typically involves the consultation of a variety of sources (manuscripts, printed editions, etc.) where the editor seeks to reconstruct the Urtext (original text, best text, etc.) of a work as "originally" created by an ancient author, while at the same time creating an apparatus criticus that records all major variants found in the sources and the reasons they chose to reconstruct the text as they did. The complicated reality of textual variants among manuscript and other witnesses of a text is one of the major reasons behind the development of modern printed critical editions, as stated by Paolo Monella:

Critical editions, i.e. editions of texts with a text-critical apparatus, respond to the necessity of representing one aspect of the complex reality of textual tradition: the textual variance. Their function is double: on the one hand, they present the different versions of a text within the context of the textual tradition; on the other hand, they try to 'extract', out of the different *texts* born by many carriers (manuscripts, *incunabula*, modern and contemporary print editions), a reconstructed *Text*, the closest possible to the 'original' one prior to its 'corruption' due to the very process of textual tradition, thus ideally recovering the *intentio auctoris* (Monella 2008).

Digital critical editions, however, offer a number of advantages over their print counterparts, according to Monella, the most important of which is the ability to better present textual variance in detail (such as by linking critical editions to the sources of variants such as transcriptions and images of manuscripts). Two other benefits of digital critical editions are first, that they allow the reader to verify and question the work of an editor, and secondly, it allows scholars to build up an "open" model of the text where the version presented by any one editor is not considered to be "the" text.

At the same time, Monella also noted that most original sources (whether manuscript or printed) in addition to including a "main text" of an ancient author also included a variety of "paratexts" that commented on it such as interlinear annotations, glosses, scholia, [220] footnotes, commentaries, and introduction. Scholia, in particular, were often considered so important and were also so vast that the scholia on a number of major authors have appeared in their own editions. [221] In order to represent the complicated nature of such sources, Monella

[220] As defined by the *Oxford Dictionary of the Classical World*, "Scholia are notes on a text, normally substantial sets of explanatory and critical notes written in the margin or between the lines of manuscripts. Many of them go back to ancient commentaries (which might fill volumes of their own). Scholia result from excerption, abbreviation, and conflation, brought about partly by readers' needs and partly by lack of space. "scholia" Oxford Dictionary of the Classical World. Ed. John Roberts. Oxford University Press, 2007. Oxford Reference Online. Oxford University Press. Tufts University. 19 May 2010 <http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t180.e1984>

[221] For example, one recently released project created by Donald Mastronarde, Professor of Classics at the University of California-Berkeley, the "Euripides Scholia Demonstration" presents a new open access digital edition of all the scholia on the plays of Euripides that were found on over twenty-nine manuscripts and printed in ten different editions. And for a distributed editing approach to scholia, a new project Scholiastae (http://www.scholiastae.org/scholia/Main_Page) has extended MediaWiki with easier word and phrase annotation in order to support individuals who wish to create their own scholia online for public domain classical texts.

proposed a model for a "document-based digital critical editions" that includes both main texts and paratexts as they appear in different individual sources:

> Such a model should include both the maintexts and the paratexts of each source, expressing explicitly the relation between single portions of each paratext and the precise portions of maintext they refer to. This implies that, rather than a traditional edition of *scholia*, it would be both an edition *of the text* and of its ancient (and modern) commentaries – and the relationships between the text and its commentaries (Monella 2008).

This model for digital critical editions then includes the need to publish each "main text" (e.g. each "reconstructed" text of an ancient author in an individual witness/source) with all of its paratexts such as scholia. Nonetheless, Monella admitted that developing a markup strategy that supports linking each paratext to the exact portion of the maintext it refers to is very difficult, and this has led to the development of a number of project specific markup strategies as well as debates over what level of "paratextuality" should be marked up in the transcriptions. Developing project specific markup is to be avoided whenever possible Monella insisted, and the raw input data (typically manuscript transcriptions in this case) should be based on existing standards so that it can be reused by other projects.

Monella ultimately recommended a fairly complicated model of transcription and markup that clearly separates the roles of transcriber and editor. Transcribers that create primary source transcriptions must confine themselves to encoding "information neutral with regards to the paratextuality levels" or else only append such information to any necessary elements with an "interpretative" attribute. An editor, who is assumed to be working interactively with a specific software tool, takes this transcription and assigns paratextuality levels to pertinent places in the transcriptions, generates an Alignment-Text of all the maintexts in the transcriptions, and stores the linking information necessary to align the maintext Alignment-Text with all of the different paratexts. The next phase involved is to create custom software that can use these objects to support dynamic and customizable access for readers to both the literary work (maintext) and its different commentary (paratexts).

The model of open source critical editions (OSCE), that has recently been described by Bodard and Garcés (2009), supported many of the conclusions reached by Monella, particularly the need to make all of the critical decisions of a scholarly editor transparent to the reader and the importance of better representing the complicated nature of primary textual sources, variants and their textual transmission. The term OSCE and its definition grow out of a 2006 meeting of scholars in the digital classics community, who met to discuss the needs of new digital critical editions in Greek and Latin:

> Our proposal is that Classical scholarship should recognize OSCEs as a deeper, richer and potentially different kind of publication from printed editions of texts, or even from digitized and open content online editions. OSCEs are more than merely the final representations of finished work; in their essence they involve the distribution of raw data, of scholarly tradition, of decision-making processes, and of the tools and applications that were used in reaching these conclusions (Bodard and Garcés, 84-85).

OSCEs are a new form of digital edition then, in that from the very beginning, they should be designed to include access to all of the raw data (page images, transcriptions), previous editions and scholarship on which this edition is based, as well as any algorithms or tools used in its creation. This argument once again reiterates the theme of the need for all digital scholarship (whether it be the creation of an archaeological reconstruction or a digital edition) to be recognized as an interpretative act.

Another significant issue addressed by Bodard and Garcés is the continuing restriction of copyright, and while they recognize that an apparatus criticus deserves copyright protection, they argue that there should be no restrictions on the public domain text of a classical author. OSCEs also highlight the changing nature of traditional scholarship and the creation of editions, by declaring that the database created or XML text behind an edition may be in many ways far more important than the scholarly prose that accompanies it:

> In the digital age, however, there is more to scholarship than simply abstract ideas expressed in elegant rhetoric and language; sometimes the most essential part of an academic work is precisely the actual words and codes used in the expression of that work….A database or XML-encoded text is not merely an abstract idea, it is itself both the scholarly expression of research and the raw data upon which that research is based, and which must form the basis of any derivative research that attempts to reproduce or refute its conclusions (Bodard and Garcés, pg. 87).

In other words, the data upon which an edition's conclusions are founded is what scholars need more than anything else, thus not only the text but all the data and the tools that produce it must be open source. The authors also observe through their brief history of the development of critical editions in the 19th and 20th century that the apparatus criticus also has a long tradition of many scholars' contributions in it and that all criticism is in the end a "communal enterprise." In fact the authors assert that text editions should only be seen as "fully critical" if all "interpretative decisions that led to the text" are made both transparent and accessible. [222]

While these requirements may seem onerous for the creation of print editions, Bodard and Garcés argue that digital editions can far more easily meet all of these demands. In addition to providing all of the data and interpretative decisions, however, they also recommend formalizing critical editions into a machine-readable form. In sum, an OSCE provides access to an open text, to the data and software used in making an edition, and to the editorial interventions made and scholarships behind the decisions.

The authors also list another major advantage of digital editions, namely the ability to get back to the materiality of actual manuscripts and move away from the "ideal" of reconstructing an Urtext, as previously discussed by (Ruhleder 1995) and (Bolter 1991), to focus instead on textual transmission. Bodard and Garcés posit that papyrologists better understand the nature of the transcription process and how creating a text is an editorial process, where there is typically not just one correct reading. Scholarship has increasingly challenged the idea that an editor can ever get back to the "original text" of an author and Bodard and Garcés stress that focus would be better paid on how to present a text with multiple manuscript witnesses to a reader in a digital environment:

> Digital editions may stimulate our critical engagement with such crucial textual debate. They may push the classic definition of the 'edition' by not only offering a presentational publication layer but also by allowing access to the underlying encoding of the repository or database beneath. Indeed, an editor need not make any authoritative decisions that supersede all alternative readings if all possibilities can be unambiguously reconstructed from the base manuscript data, although most would in practice probably want to privilege their favoured readings in some way. The critical edition, with sources fully incorporated, would potentially provide an interactive resource that assists the user in creating virtual research environments (Bodard and Garcés, pg. 96).

Thus the authors hope that digital or virtual research environments will support the creation of "ideal" digital editions where the editor does not have to decide on a "best text" since all editorial decisions could be linked to their base data (e.g., manuscript images, diplomatic transcriptions). The creation of such "ideal" digital editions, however, they also urge must be a collaborative enterprise, where all modifications and changes are made explicit, are attributed to individuals, and are both citable and permanent. Bodard and Garcés thus conclude that future research should examine what methodologies and technology are necessary to make this vision a reality. A key part of this research will be to explore the relationship between OSCEs and the materials found in massive digital collections and million book libraries with little or no markup. OSCEs and other small curated collections, Bodard and Garcés insist, can be used as training data to enrich larger collections, a theme which we will return to in the discussion of classics and cyberinfrastructure.

One project that embodies some of the principals discussed here is the Homer Multitext Project (HMT), [223] defined by Blackwell and Smith (2009) as an "effort to bring together a comprehensive record of the Homeric tradition in a digital library." The website for the HMT is hosted by the CHS and offers free access to a library of text transcriptions and images of Homeric manuscripts, with its major component being digital images of the tenth century manuscript of the *Iliad* known as the Venetus A from the Marciana Library in Venice. According to the website:

> This manuscript, the oldest and best, on which all modern editions are ultimately based, contains in its marginal commentaries a wealth of information about the history of the text. These commentaries in the margin, or scholia, derive mainly from the work of scholars at the Library at Alexandria in Egypt during the Hellenistic and Roman eras. It has been a central goal of the project to obtain and publish high-

---

[222] Similar arguments have been by Henriette Roued-Cunliffe in regards to creating a reading support system for papyrologists that models and records their interpretative processes as they create an edition of a text: "It is very important that this evaluation of the evidence for and against the different readings is conducted. However, the commentary only presents the conclusions of this exercise. It would be a great aid, both for editors as they go through this process, and also for future editors of the same text, if it were possible to present this evaluation for each character and word in a structured format" (Roued-Cunliffe 2010)

[223] http://chs.harvard.edu/wa/pageR?tn=ArticleWrapper&bdc=12&mn=1169. While the discussion here will focus largely on the technical architecture and innovative approach to digital editions of the Homeric Multitext project, a long-term view of the scholarly potential and future of the project has recently been offered by (Nagy 2010).

resolution digital images of the manuscript, together with an electronic edition of the Greek text of the scholia along with an English translation. [224]

In order to represent such a complicated manuscript, Blackwell and Smith reported that "the HMT has focused not on building a single-purpose application to support a particular theoretical approach, but on defining a long-term generic digital library expressly intended to encourage reuse of its contents, services, and tools."

As was earlier observed by Bodard and Garcés, the 20th century began a movement from scholarship based on manuscripts towards the creation of critical editions. Blackwell and Smith list some of the most prominent Homeric editions including editions of the scholia, and argue that they all suffer from the same major flaw, they are works based on selection where the main goal of the editor was to present a unified text that represented their best judgment, with the result being that many scholia or variants were often excluded. Recent changes in Homeric textual criticism, however, and the development of digital technology have allowed these questions of textual variation and types of commentary to be revisited:

> The very existence of variation in the text has become a matter of historical interest (rather than a problem to be removed). The precise relationship between text and commentary, as expressed on the pages of individual manuscripts, hold promise to shed light on the tradition that preserved these texts, the nature of the texts in antiquity, and therefore their fundamental nature…. The best scholarly environment for addressing these questions would be a digital library of facsimiles and accompanying diplomatic editions. This library should also be supplemented by other texts of related interest such as non Homeric texts that include relevant comments and quotations and other collections of data and indices. Thus our focus on both collection of data and on building a scalable, technologically agnostic, infrastructure for publishing collections of data, images, texts, and extensions to these types (Blackwell and Smith 2009).

Thus the HMT project stresses the importance of providing access to the original data used to create critical editions such as manuscript images, the need for diplomatic transcriptions that can be reused, and a related body of textual and other material that helps to more firmly place these manuscripts in their historical context and better supports the exploration of their textual transmission. The HMT makes extensive use of the CTS protocol and its related Reference Indexing service, both of which exist as "JavaServlets and Python applications running in the Google AppEngine." The HMT has also developed its own web-based interface to their library called PanDect. [225]

The HMT's inclusion of scholia (and work done by the related Homer and the Papyri project) [226] also highlights the fact that manuscripts included many "paratexts" as previously described by Monella (2008). In fact, Blackwell and Smith note that the Venetus A, like many Byzantine codices, contained many discrete texts, including a copy of Proclus's *Chrestomathy*, the *Iliad*, summaries of books of the *Iliad*, four different scholiastic texts, and later notes. In their system of text linking and retrieval through abstract citation, all of these contents are described as separate texts and the CTS protocol is used to refer to the structure of each text, and indices are then created that associate these texts with digital images of the folio sides that make up the manuscript. [227] Ultimately, Blackwell and Smith convincingly argue that this approach to primary sources that favors diplomatic or facsimile editions and simple indexing over complicated markup will support the greatest possible amount of reuse for the data, positing that it "requires less knowledge to integrate texts with simple markup and simple, documented indices, than to disaggregate an elaborately marked up texts that embeds links to other digital objects."

While all primary sources might benefit from such an infrastructure, the Homeric poems in particular require such a special infrastructure, according to Dué and Ebbott (2009), largely due to the complicated oral performance tradition that created the poems. Traditional printed editions of the poems with a main text and an apparatus that records all alternative interpretations they argue creates the misleading impression that there is "one" correct text and then there is everything else. Dué and Ebbott also criticize the fact that the critical apparatus can only be deciphered by a specialist audience with years of training. The digital medium can be used to better represent the nature of the Homeric texts, however, for as they contend:

---

[224] http://chs.harvard.edu/wa/pageR?tn=ArticleWrapper&bdc=12&mn=1381
[225] http://pandect.sourceforge.net/
[226] http://chs.harvard.edu/wb/93/wo/1DSibj9gCANI20WSihcJNM/0.0.0.0.19.1.7.15.1.1.0.1.2.0.4.1.3.3.1
[227] For extensive detail on these text structures, data models and more technical implementation details, please see (Smith 2010)

> The Homeric epics were composed again and again in performance: the digital medium, which can more readily handle multiple texts, is therefore eminently suitable for a critical edition of Homeric poetry—indeed, the fullest realization of a critical edition of Homer may require a digital medium (Dué and Ebbott 2009).

According to their argument that the Homeric poems were composed anew in every performance, there is no *single* original author's composition or text to attempt to reconstruct. They put forward that the variations found in different manuscript witnesses are not necessarily copying "errors" and that in fact the traditional term "variants" as used by textual critics is not appropriate for the compositional process used to create the poems. Dué and Ebbott assert that the digital medium can support a superior form of textual criticism for these epics:

> Textual criticism as practiced is predicated on selection and "correction" as it creates the fiction of a singular text. The digital criticism we are proposing for the Homer Multitext maintains the integrity of each witness to allow for continual and dynamic comparison, better reflecting the multiplicity of the textual record and of the oral tradition in which these epics were created (Dué and Ebbott, 2009).

Instead of the term variants, they coin the term "performance multiforms" to describe the variations found in the manuscript witnesses. Standard printed critical editions could never reflect the complexity of such multiforms, Dué and Ebbott submit, but they insist that a digital "edition" such as the HMT supports the representation of various manuscript witnesses and can more clearly indicate where variations occur, since no "main text" must be selected for presentation as on the printed page.

Yet another advantage of the multitextual approach, Dué and Ebbott reveal, is that it can be far more explicit about the many channels of textual transmission. For example, many quotations of Homer in Plato and the Attic orators as well as fragmentary papyri are often quite different from the medieval texts that served as the basis for all modern printed editions of Homer. A multitext approach allows each of these channels to be placed in a historical or cultural framework that can help the reader to better understand how they vary, rather than an apparatus that often obfuscates these differences. Nonetheless, Dué and Ebbott acknowledge that building a multitext that moves from a "static perception to a dynamic presentation" and attempts to present all manuscript witnesses to a reader without an intervening apparatus faces a number of technical challenges that are still being worked out, including how to highlight multiforms so they are easy to find and compare, and how to display hexameter lines (the unit of composition in the Homeric epic) as parts of whole texts rather than just pointing out the differences (as in an apparatus). While these issues are still being worked out, the authors conclude that three main principles drive their ongoing work: collaboration, open access and interoperability.

Similar criticism of modern critical editions and their inability to accurately represent the manuscript tradition of texts has also been offered by Stephen Nichols. Nichols stated that the modern editorial practice of attempting to faithfully reconstruct a text as the original author intended it has little to do with the "reality of medieval literary practice" and is instead an "artefact of analogue scholarship" where the limitations of the printed page required editors to choose a base manuscript to transcribe and to banish all interesting variants from other manuscripts to the apparatus (Nichols 2009). He also voiced that there was very little interest in providing access to original manuscripts, as many scholars considered the scribes who produced them to have introduced both copying errors and their own thoughts and thus to have "corrupted" the original text of the author. The advent of digital technology, however, Nichols concluded had produced new opportunities for studying literary production:

> The Internet has altered the equation by making possible the study of literary works in their original configurations. We can now understand that manuscripts designed and produced by scribes and artists—often long after the death of the original poet—have a life of their own. It was not that scribes were 'incapable' of copying texts word-for-word, but rather that this was not what their culture demanded of them. This is but one of the reasons why the story of medieval manuscripts is both so fascinating, and so very different from the one we are accustomed to hearing. But it requires rethinking concepts as fundamental as authorship, for example. Confronted with over 150 versions of the work, no two quite alike, what becomes of the concept of authorial control? And how can one assert with certainty which of the 150 or so versions is the 'correct' one, or even whether such a concept even makes sense in a pre-print culture (Nichols 2009).

Thus the digitization of manuscripts and the creation of digital critical editions has not only provided new opportunities for textual criticism, but might even be viewed as enabling a type of criticism that better respects the traditions of the texts or objects of analysis themselves.

While Monella (2008), Bodard and Garcés (2009), and Dué and Ebbott (2009) focused largely on the utility of digital editions for philological study and textual criticism, Notis Toufexis has also recently argued that digital editions are central to the work of historical linguistics as well. As he explains, historical linguistics "examines and evaluates the appearance of new—that is changed—linguistic forms next to old (unchanged) ones in the same text or in texts of the same date and/or geographical evidence (Toufexis 2010, pg. 111). Similar to Stephen Nichols, Toufexis criticized modern critical editions for creating a far simpler linguistic picture than was actually the case within medieval manuscripts. He described how scribes might have unconsciously used newer forms of language and not copied the old forms found in a manuscript or how they might have made specific decisions to use older forms as a stylistic choice to elevate the register of the text. Thus the inclusion of all text variants in the apparatus criticus is necessary not just for philologists but also for historical linguistics who wish to examine how linguistic features have changed across historical corpora. Toufexis argued that digital editions could thus solve the problems of both philologists and historical linguists:

> A technology-based approach can help us resolve this conflict: in a digital environment 'economy of space' is no longer an issue. By lifting the constraints of printed editions, a digital edition can serve the needs of both philologists and historical linguistics (or for that matter any other scholar who has an interest in approaching ancient texts). A 'plural' representation of ancient texts in digital form, especially those transmitted in 'fluid' form, is today a perfectly viable alternative to a printed edition. Only a few years ago such a digital endeavor seemed technologically impossible or something reserved for the very few computer-literate editors (Toufexis 2010, pg. 114-115).

Toufexis hoped that even if critical editors could not be convinced to change the way they edited texts, that most of the problems of critical editions could at least be ameliorated by being transposed to a digital medium, because digital editions could make editorial choices transparent by linking the apparatus criticus to the electronic text and could be accompanied by digital images of manuscript witnesses. Digital editions, Toufexis argued, were ultimately far better for readers as well, because "a pluralistic digital edition encourages readers to approach all transmitted texts equally, even if one text is highlighted among the many texts included in the edition" (Toufexis 2010, pp. 117-118).

New Models of Collaboration, Tools & Frameworks for Digital Editions

Digital tools create new opportunities for textual editing and the creation of digital editions, and one key area of opportunity is their ability to support new types of collaboration. Tobias Blanke has recently suggested that "traditional humanities activities such as the creation of critical editions could benefit from the collaboration in research enabled by new infrastructures that e-Science promises to deliver" (Blanke 2010). Indeed, Peter Robinson has argued that the single greatest shift in editing practice brought about by the digital world is "that it is creating new models of collaboration: it changes who we collaborate with, how we collaborate, and what we mean by collaboration" (Robinson 2010).

Robinson convincingly argues that the first digital editions did not challenge the traditional editorial model, where a single editor frequently gathered materials and made all final editing decisions even if they had a number of partners in terms of publishing, the final product was usually under the control of one person. In the digital world, Robinson proposes a new model is possible where for example, libraries can put up images of manuscripts, various scholars/students or experts can make transcriptions that link to these images, other scholars can collate these transcriptions and publish editions online linking to both the transcriptions and images, yet more scholars can analyze these collations and create an apparatus or commentaries, and other scholars can then link to these commentaries. All of these activities can occur independently or together.

While Robinson grants that more traditional single-editor controlled editions can be made in the digital world, such editions are far too expensive, he posits, particularly since one can't just present samples online but needs to provide access to all images and transcriptions of the text. The single editor model he believes will also lead inevitably to the sole creation of a limited number of digital editions of major works by major authors. Robinson reported that there was a scholarly backlash against the creation of such high profile and expensive digital editions in the last few years, where the divide was largely between those with access to expensive tools and those without. He concludes that this backlash directly contributed to the closing of the Arts and Humanities Data Service (AHDS).

In this new digital world with an endless amount of editing to be done, Robinson urges humanists to actively guide the building of the necessary infrastructure by providing tools, examples of good practice, and key parts of their own editions, since after all, it is in their own best interest as well to have a say in any infrastructure designed for them. He stresses, however, that, "there is not and will not be a Wikipedia for editions, nor indeed will there ever be any one tool or software environment which does everything for everyone. What there might be is a set of tools and resources, built on agreed naming conventions…" (Robinson 2010). He thus argues for basic standards and naming conventions but against any massive universal infrastructure. Robinson defines what he is calling for as "distributed, dynamic and collaborative editions" (Robinson 2009) and argues that:

> The concept is not that there is a single system, a single set of software tools, which everybody uses. Instead, across the web we have a federation of separate but co-operating resources, all within different systems, but all interlinked so that to any user anywhere it appears as if they were all on the one server (Robinson 2009).

In fact, Robinson expressed frustration at the fact that most funding agencies were obsessed with what he considered to be "Grand Single Solutions." He saw little utility in projects such as SEASR or Bamboo, and summed up his opinion thusly: "Let me say this clearly, as most scholars seem afraid to say it: projects like these are vast wastes of time, effort and money" (Robinson 2009). He argued that the future did not lie with massive single purpose infrastructure, but instead with projects such as Interedition [228] and the Virtual Manuscript Room, [229] that are not only seeking ways to both create more resources such as manuscript images online but also to link disparate parts (images, tools, transcriptions) that are already online together, rather than trying to force them all into one new infrastructure.

The Interedition Project is seeking to create an "interoperable supranational infrastructure for digital editions" across Europe that will promote interoperability of the tools and methodologies used in the field of digital textual editing. While this project is still is in its early stages, it is currently working on a tool called CollateX, [230]"a java library for collating textual sources" that is the latest version of the tool Collate. Collate was originally created by Peter Robinson to support the collation of multiple versions of an electronic text in order to create scholarly editions, and its functionalities are being enhanced in CollateX (Robinson 2009).

In his discussion of Collate, Robinson emphasized that developers of scholarly collation tools should recognize two essential facts, the first and foremost of which is that "scholarly collation is not Diff" and so any attempt to build a collation program that meets all scholars' needs will meet with failure. He lists two major reasons for this, first, while automated programs can easily identify differences these differences are not necessarily variants, and second, teaching a machine to determine the best way to present "any given sequence of variants, at any particular moment" would be a monumental task. One of the features Robinson argued that gave Collate a reasonable measure of uptake was the fact that it "allows the scholar to fix the collation exactly as he or she wants" (Robinson 2009).

The second essential fact that Robinson felt developers should recognize was that "collation is more than visualization." While many collation programs can beautifully show variation, Robinson acknowledged, they can present it in one format only, so he designed Collated differently:

> Again, one thing I did right with Collate, right back in the very beginning, was that I did not design the program so it could produce just one output. I designed it so that you could generate what you might call an intelligent output. Essentially, this distinguished all the different components of an apparatus - the lemma, the variant, the witness sigil, and much more – and allowed you to specify what might appear before and after each component, and in what order the various components might appear (Robinson 2009).

Collate also allowed scholars to output the apparatus in forms ready for processing by various analysis programs. At the same time, Collate has a number of issues that have required the development of CollateX, including difficulties handling transpositions and even more critically the inability to support collaborative work.

---

[228] http://www.interedition.eu/
[229] http://vmr.bham.ac.uk/
[230] https://launchpad.net/collatex

A number of smaller research projects are also attempting to build environments where humanists can work together collaboratively on texts and digital editions. The Humanities Research Infrastructure and Tools (HRIT) [231] project based at the Center for Textual Studies and Digital Humanities at Loyola University Chicago is currently working on a tool called e-Carrel, that will create a secure and distributed infrastructure for access, preservation and re-use of humanities texts. The e-Carrel tool will also support the use of a collaborative annotation tool and standoff markup. In particular, the creators of e-Carrel want to support interoperability with other humanities text projects, and promote collaborative work on a series of "core texts" that will likely exist as multi-versioned documents (Thiruvathukal et al. 2009).

In addition to collaborative tools, a number of other tools exist to support the creation of digital editions. One well-established "tool" for creating digital editions is "Edition Production & Presentation Technology" (EPPT) [232] a set of XML tools that have been designed to assist editors in creating image based electronic editions. The EPPT is a stand-alone application that editors can install on their own computers and that supports more effective "image based encoding" or where users link descriptive markup (such as a TEI transcription of a manuscript) to material evidence (an image of that manuscript) through XML. Templates are automatically generated from the data of individual projects so that scholars and students need little training in TEI/XML to get started. EPPT has two basic tools for integrating images and text, ImagText (with OverLay) and xTagger and enables very precise linking of both full images and image sections with structural and descriptive metadata. To create a basic image based edition a user simply needs images, corresponding plain text and a DTD. Consequently EPPT can be used to create image-based editions using images and data available in different archives, and can also be used by scholars to "prepare, collate and search" variant manuscript versions of texts. A number of projects have already made use of this tool including the Roman de la Rose and Electronic Boethius. [233]

A related project in terms of desired functionality is TILE (Text Image Linking Environment) [234], a collaborative project of the Maryland Institute for Technology in the Humanities, [235] the Digital Humanities Observatory (DHO), [236] and Indiana University Bloomington. This two year project is seeking to "develop a new web-based, modular, collaborative image markup tool for both manual and semi-automated linking between encoded text and image of text, and image annotation." Doug Reside of this project recently outlined on the TILE blog a "four layer model for image-based editions" that was designed to address long-term preservation issues and clearly outline responsibilities for digital librarians and scholars (Reside 2010).

The first level involves the digitization of source materials, particularly their long-term curation and distribution in open formats with the use of regular and progressive naming systems. Reside makes a useful suggestion that granting agencies should consider requiring content providers to maintain stable URIs for at least ten to fifteen years for all digital objects. The second level involves metadata creation, and Reside argues that all metadata external to the file itself (e.g. descriptive rather than technical metadata) belongs at this level. He also proposes that such metadata might best be created by institutions or individuals that did not create the digital files:

> While the impulse towards quality assurance and thorough work is laudable, a perfectionist policy that delays publication of preliminary work is better suited for immutable print media than an extensible digital archive. In our model, content providers need not wait to provide content until it has been processed and catalogued (Reside 2010).

By opening up the task of cataloging and resource description to a larger audience, Reside notes far more content can get online quickly and be available for reuse. Separating metadata and content would also allow multiple transcriptions or metadata to point to the same item's URI.

---

[231] http://text.etl.luc.edu/HRIT_CaTT/overview.php
[232] http://www.eppt.org/eppt/. For more on the technical approach behind the EPPT and using XML for "image-based electronic editions" see Dekhytar et al. (2005).
[233] http://beowulf.engl.uky.edu/~kiernan/eBoethius/inlad.htm
[234] http://mith.info/tile/
[235] http://mith.info/
[236] http://dho.ie/

The third level of the TILE model involves the interface layer, an often ignored feature in the move to get open content available online. While Reside grants that more transcriptions and files in open repositories is a useful first step, many humanities scholars need interfaces to do more than access on file at a time. He also recognized that while SEASR is trying to create a sustainable model for interoperable digital humanities tools, their work has not yet met with wide-scale adoption. At this most critical layer, Reside outlines the TILE approach:

> We propose a code framework for web-based editions, first implemented in JavaScript using the popular jQuery library, but adaptable to other languages when the prevalent winds of web development change. An instance of this framework is composed of a manifest file (probably in XML or JSON [237] format) that identifies the locations of the relevant content and any associated metadata and a core file (similar to, but considerably leaner than, the core jQuery.js file at the heart of the popular JavaScript library) with a system of "hooks" onto which developers might hang **widgets** they develop for their own editions. A **widget**, in this context, is a program with limited functionality that provides well-defined responses to specific input (Reside 2009).

This model thus includes a *manifest* file that contains all content locations and associated metadata and a *core file* or base text that can be used by different developers to create their own digital editions utilizing their own tools or "widgets." Widgets should only depend on the core files, Reside argues, not each other, and ideally could be shared between different scholars. Reside admits that basically they are proposing the development of a "content management system" (CMS) for managing "multimedia scholarly editions" even though the market is currently crowded with different CMS options, but according to Reside none of the currently available options quite meets the needs of scholarly editions.

The fourth and final layer of the TILE model involves user generated data layers, and Reside considers this to possibly be the most "volatile data in current digital humanities scholarship." Furthermore, the open nature of many sites makes it hard to distinguish contributions from the inexperienced versus expert scholars. Thus while their framework argues for the "development of repositories of user-generated content," since all content contributed by users cannot be permanently stored, they suggest "sandbox" databases where only the best user-generated content is selected for inclusion and publication.

One partner in the TILE project, the DHO has also conducted some independent research into developing a framework for scholarly editions (Schreibman 2009). Schreibman offered similar criticisms to those of Robinson and Reside, stating that not only were many early digital editions typically one-off productions where the content was tightly integrated with the chosen software, but that:

> We also don't, as a scholarly editing community, have agreed upon formats, protocols, and methodologies for digital scholarly editing and editions. Moreover, many of the more mature first-generation digital projects creating online editions from print sources have more in common with digital library projects—i.e. editions created with a light editorial hand, minimally encoded and with little more contextualization than their print counterparts (Schreibman 2009).

In order to address some of these issues the DHO held a one day symposium in 2009 on the issue of digital editions that was then followed up by a week long spring scholarly editing school to determine "a set of protocols, methodologies, rights management and technical procedures to create a shared infrastructure for digital scholarly editions in Ireland." They also plan to follow relevant developments from TextGrid and Interedition, so that the infrastructure and tools developed in Ireland can link up with these other national and international projects.

The Challenges of Text Alignment & Text Variants

As illustrated by the above discussion of digital editions, any infrastructure developed for digital classics and for the creation of digital editions will need to consider the challenges of both text alignment and textual variants. The research literature on this topic is extensive and this subsection will briefly consider two recent state-of-the-art approaches to deal with these issues. [238]

While the introduction to this review illustrated that there are a large number of digital corpora available in both Greek and Latin, Federico Boschetti (2007) has criticized the fact that although these corpora are typically based

[237] JSON, or JavaScript Object Notation is a "lightweight data-interchange format" that is based on the JavaScript programming language but is also a "text format that is completely language independent." (http://www.json.org/)

[238] For a thorough bibliography of over 50 papers in this area see the list of references in (Schmidt and Colomb 2009).

on authoritative editions they also provide no access to the apparatus criticus. [239] When using a literary corpus such as the TLG, Boschetti reminds his readers that they must remember they are dealing with the text of an author that has been created by editorial choices. This makes it particularly difficult to study linguistic or stylistic phenomenon for without access to the apparatus criticus it is impossible to know what variants an editor may have suppressed. This can render digital corpora useless for philologists as Boschetti explained:

> Philologists use digital corpora but they must verify results on printed editions, in order to evaluate if the text retrieved is attested in every manuscript, only in the *codex optimus*, in an error prone family of manuscripts, in a *scholium*, in the indirect tradition or if it is conjectured by a modern scholar. In short, the text of the reference edition has no scientific value without the apparatus…(Boschetti 2007).

He also noted, however, that two exceptions, to this phenomenon are the Homer Multitext Project and Musisque Deoque, [240] both of which are seeking to enrich the corpora they create with variants and conjectures.

Boschetti articulated that there were two basic methods to add apparati critici to digital critical editions. The first method was based on the automatic collation of diplomatic editions, where digital diplomatic editions are defined as "complete transcriptions of single manuscripts" with encoded information about text layout and position (typically encoded in TEI-XML). In agreement with Monella (2008), Boschetti commented that one of the most useful features of markup such as TEI is that it makes it "possible to separate the actual text of the manuscript from its interpretations." This method was particularly useful Boschetti argued for texts with a limited number of manuscripts. The second method involves the manual filling of forms by manual operators, an approach utilized by Musisque Deoque, and according to Boschetti, is "useful if the aim is the acquisition of large amounts of apparatus' information, on many texts of different authors." Both of these approaches also have shortcomings, Boschetti pointed out, for the collation of diplomatic editions must be integrated with other techniques, and manual form filling is subject to human error.

Either approach would also be unfeasible for an author like Aeschylus with an extensive body of secondary analysis and large numbers of conjectures registered in various commentaries and reviews. Boschetti thus proposed a third approach combining automatic collation and information extraction:

> The automatic parsing of apparatuses and repertories, in addition to the automatic collation for a group of relevant diplomatic transcriptions, should be an acceptable trade-off. Subjective choices by operators in this case is limited to the correction phases. This third approach has a double goal: on one hand it aims to parse automatically existing critical apparatuses and repertories of conjectures of Aeschylus and on the other hand it aims to discover heuristics useful for any collection of variants and/or conjectures with a similar structure (Boschetti 2007).

Boschetti designed a complete methodology that began with the automatic collation of three reference editions for Aeschylus so that there would be a unified reference edition on which to map the apparatuses and repertories of conjectures. The next step was to conduct a manual survey of various apparatuses in order to identify typical structures (e.g. verse number, reading to substitute a word in text, manuscript and scholar). After identifying references to verses, Boschetti developed a typology of readings and sources for the information in the apparatus. He noted the most frequent case involved readings where an orthographic or morphological variant would "substitute a single word in the reference edition." The most common other operations were deletion, addition and transposition of text. In terms of sources, they were typically "one or more manuscripts for variants" and "one or more scholars for conjectures" occasionally followed by accurate bibliographical information. One major difficulty, Boschetti noted, was that the same manuscript or author could often be abbreviated differently in different apparatuses. In the system Boschetti developed, names had to "match items of a table that contains the canonical form of the name, abbreviations, orthographical variants and possible declinations"

The next major step was to develop a set of heuristics to be used in automatically parsing the different apparatuses. Each item in the apparatus was separated by a new line and then all items were tokenized into one of the following categories: verse number, Greek word, Greek punctuation mark, metrical sign, Latin word,

---

[239] This criticism has also been made by (Ruhleder 1995) and (Stewart, Crane and Babeu 2007).

[240] Musisque Deoque is "a digital archive of Latin poetry, from its origins to the Italian Renaissance" that was established in 2005 and is creating a Latin poetry database that is "supplemented and updated with critical apparatus and exegetical equipments." http://www.mqdq.it/mqdq/home.jsp?lingua=en.

Latin punctuation mark, scholar name, manuscript abridgement, and bibliographic reference. All scholars' names, manuscript abridgments and bibliographic references were compared with information from the tables created in the previous step. The rest of the tokens were then aggregated to identify verse references, readings and sources. The final step involved in this process was the use of an alignment algorithm to parse text substitutions "in order to map the readings on the exact position of the verse in the reference edition." Boschetti revealed that about 90 percent of readings found in apparatuses were *substitutions*, or chunks of text that should replace one or more lines in a reference edition. His algorithm utilized the concept of "edit distance" [241] to align readings from the apparatus with the portion of text in the reference edition where the edit distance was lowest. Boschetti also chose to use a "brute force" combinatorial algorithm that "reconstructs all the combinations of adjacent words in the reference text (capitalised and without spaces) and it compares them with the reading and its permutations." One current limitation of his work Boschetti reported was that the current system is only applied on "items constituted by Greek sequences, immediately followed by source" and excludes those cases where items included Latin language explanations of textual operations to perform or a judgment.

To test his system Boschetti calculated its performance against 56 verses of Wecklein's edition of Aeschylus' *Persae* and evaluated it by hand. For processed items (excluding items with Latin predicates), 88 % of conjectures were mapped onto the reference text correctly, and 77% of conjectures were mapped correctly in the total collection. The work conducted by Boschetti illustrates that even while an automated system did require a fair amount of preliminary manual analysis, the heuristics and algorithm that were created provided encouraging results that deserve further exploration.

Recent work by Schmidt and Colomb (2009) has taken a different approach to the challenge of textual variation, one that also addresses related issues with overlapping hierarchies in markup. According to Schmidt and Colomb there are two basic forms of textual variation, that found in multiple copies of a work such as in the case of multiple manuscripts, and that which arises from physical alterations introduced by an author or copyist in a single manuscript. Both early printed books and handwritten medieval manuscripts often have high levels of variation and the techniques of textual criticism grew up around the desire to create a single, definitive text. Despite the fact that the digital environment provided new possibilities for representing multiple versions of a text, significant disagreement among textual editors continued as Schmidt and Colomb related:

> With the arrival of the digital medium the old arguments gradually gave way to the realisation that multiple versions could now coexist within the same text….This raised the prospect of a single model of variation that might at last unite the various strands of text-critical theory. However, so far no generally accepted technique of how to achieve this has been developed. This failure perhaps underlies the commonly held belief among humanists that any computational model of a text is necessarily temporary, subjective and imperfect (Schmidt and Colomb 2009).

Additionally, Schmidt and Colomb proposed that the lack of an "accurate model of textual variation" and how to implement it in a digital world continued to frustrate many humanists.

A related problem is that of *overlapping hierarchies* or when different markup structures overlap in a text (e.g. generic structural markup, linguistic markup, literary markup). Markup is said to overlap in that "the tags in one perspective are not always well formed with respect to tags in another" (e.g. as in well-formed XML). Schmidt and Colomb propose that the term overlapping hierarchies is essentially incorrect: "Firstly, not all overlap is between competing hierarchies, and secondly what is meant by the term 'hierarchy' is actually 'trees', that is a specific kind of hierarchy in which each node, except for the root, has only one parent." They put forward that although there have been over 50 papers dealing with this topic one fundamental weakness in all of the proposed approaches was that they all offer solutions to problematic markup by using *markup* itself. In addition, they go further and assert that all cases of *overlapping hierarchies* are also cases of *textual variation*, even if the reverse is not always true. "The overlapping hierarchies problem, then, boils down to variation in the metadata," Schmidt and Colomb declared, "It is entirely subsumed by the textual variation problem because textual variation is variation in the entire text, not only in the markup" (Schmidt and Colomb 2009). They thus concluded that textual variation was the problem that needed solving.

---

[241] Edit distance has been defined as a "string distance" or as the number of operations required to transform one string into another (with typical allowable operations including the insertion, deletion, or substitution of a single character) http://en.wikipedia.org/wiki/Edit_distance

Schmidt and Colomb state that neither version control systems nor multiple sequence alignment (inspired by bioinformatics) can adequately address the problem of text variants and instead propose modeling text variation as either a "minimally redundant directed graph" or as an "ordered list of pairs" where each pair contains a "set of versions and a fragment of text or data." The greatest challenge with variant graphs they explained is how to process them efficiently, and the minimum number of functions that users would need to be available included: reading a single version of a text, searching a multi-version text, comparing two versions of a text, determining what was a variant of what else, creating and editing a variant graph, and separation of content and variation. The proposed solution outlined by Schmidt (2010) is the multi-version document format (MVD):

The Multi-Version Document or MVD model represents all the versions of a work, whether they arise from corrections to a text or from the copying of one original text into several variant versions, or some combination of the two, as four atomic operations: insertion, deletion, substitution, and transposition…. An MVD can be represented as a directed graph, with one start node and one end-node… Alternatively it can be serialized as a list of paired values, each consisting of a fragment of text and a set of versions to which that fragment belongs. As the number of versions increases, the number of fragments increases, their size decreases, and the size of their version-sets increases. This provides a good scalability as it trades off complexity for size, something that modern computers are very good at handling. By following a path from the start-node to the end-node any version can be recovered. When reading the list form of the graph, fragments not belonging to the desired version are merely skipped over (Schmidt 2010)

In addition, Schmidt lists a number of benefits of the MVD format for humanists including 1) it supports the automatic computation of insertions, deletions, variants and transpositions between a set of versions, 2) MVDs are content format-agnostic about individual versions so they can be used with any generalized markup or plain text, 3) a MVD is "not a collection of files" and instead stores "only the differences between all the versions of a work as one digital entity and interrelates them" (Schmidt 2010), 4) since the MVD stores the overlapping structures of a set of versions the markup of individual texts can be much simpler, and 5) "an MVD is the format of an application not a standard." They suggest that MVD documents should be stored in a binary format, particularly if the content of each text is in XML. In their current work, they have created a MultiVersion wiki tool where scholars can work on cultural heritage texts that exist in multiple versions.

Epigraphy

Overview: Epigraphy Databases, Digital Epigraphy and EpiDoc

Epigraphy has been defined as the study of "inscriptions or epigraphs engraved into durable materials (e.g. stone)" (Bauer et al. 2008) and this digitally advanced discipline is well represented online by numerous projects as well as a relatively mature encoding standard EpiDoc. [242] According to the Corpus Inscription Latinarum (CIL) project, "Inscriptions, as direct evidence from the ancient world, are among the most important sources for investigating Roman history and everyday life in all their aspects." [243] Similarly, Bodard (2008) has offered further explanation of the importance of inscriptions for classical scholarship:

Inscriptions, ancient texts inscribed on stone or other durable materials, are an important source of access to various ancient societies, and particularly the worlds of ancient Greece and Rome. These texts survive in large numbers, and are widely used by historians as one of the primary sources of direct evidence on the history, language, rituals, and practices of the ancient world. Words inscribed on stone, a skilful and expensive process, may tend to be élite texts…(Bodard 2008).

Nonetheless, Bodard also stated that in addition to official documents there are many other types of inscriptions such as gravestones and curse tablets that give insight into the everyday life of ordinary people.

A recent overview of the state-of-the-art in digital epigraphy [244] and the future of epigraphy as a discipline was given by Cayless et al. (2009). They stated that while the majority of epigraphic publications are currently still published only in print that by 2017 this will have changed. The discipline of epigraphy grew greatly during the 18th and 19th century, Cayless et al. observed, both as a standard education for gentleman in Latin and Greek and travel in the eastern Mediterranean increased. Many inscriptions were transcribed by non-classical scholars, but a scientific approach for transcribing gradually developed as did standards for publication, albeit in a rather

[242] http://epidoc.sourceforge.net/
[243] http://cil.bbaw.de/cil_en/dateien/forschung.html
[244] For an overview of the state-of-the-art of digital research methods specifically for Latin epigraphy, see (Feraudi- Gruénais 2010).

haphazard manner. In the early 1930s, a set of publishing protocols called the Leiden conventions (van Groningen 1932) were agreed upon, conventions that have been discussed and updated ever since according to Cayless et al. (2009). The Leiden conventions have been described as "a type of semantic encoding, which consists of various brackets, underdots and other markings relating to missing or broken characters, uncertainty, additions and corrections made by the editor of an ancient text" (Roued 2009). Despite the creation of these conventions, Roued (2009) noted that editions published before 1931 used varying conventions and even after the creation of Leiden, not all parts of the conventions were applied evenly.

One major issue with standard print publication in epigraphy Cayless et al. observed was that it "tended to emphasize the role of epigraphy within archaeology and history, and to distance it from the study of text and language." Bodard (2008) has also emphasized this unique feature of inscriptions:

> The texts themselves are an awkward category, neither poetry, history, or philosophy, nor even in the same category as literature preserved by the direct manuscript tradition, but documentary texts with very little beauty or elegance of language. The objects on which the texts are inscribed, the stelae, statues, wall panels, tablets, and grave monuments, are studied by archaeologists and art historians for whom the written texts are little more than a footnote, if not an inconvenience. This fact has tended to keep inscriptions in an academic limbo—not quite literary text and not quite archaeological object (Bodard 2008).

In fact, Bodard claimed that electronic publication supports an entire reappraisal of inscriptions, and that text encoding and subject based markup, in particular, increase the ability to deal with inscriptions as both texts and archaeological objects.

In order to delineate the future of digital epigraphy, Cayless et al. (2009) referred to John Unsworth's list of scholarly primitives (Unsworth 2000), "discovery, annotation, comparing, referring, sampling, illustrating, and representing" and then used it as a framework for analyzing how well epigraphy databases addressed these needs. They argued that epigraphy databases have been greatly successful in supporting the task of *discovery*, and that providing the ability to search across texts has been one of the major goals behind most digital epigraphy projects. In addition, any project published online can also be searched at least by Google. Indeed, as the survey of projects below will illustrate, the majority of digital epigraphy projects are database driven. Cayless et al. recommended however, that the standard approach taken by most epigraphy projects fails to address the other scholarly primitives, and a different type of digital representation is thus necessary.

To support this assertion, Cayless et al. briefly reviewed EDH, EAGLE, and several other digital epigraphy projects, and suggested that they represented standard approaches to digitally representing inscriptions and related data. They also pointed out that growing massive digitization projects such as Google Books and the Internet Archive have also scanned a number of public domain editions of inscriptions (though Google Books sometimes restricts access to some of these texts without explanation, particularly to users outside of the United States). The standard approach of most databases as described by Cayless et al. directly transfers a Leiden encoded inscription to digital form with only some adjustments. In contrast, they advocate the use of EpiDoc, a TEI XML standard created by Tom Elliott for encoding inscriptions. Although originally conceived of as a common data interchange format, Cayless et al. reported that through a number of projects and workshops:

> …EpiDoc has grown and matured. Its scope has expanded beyond (though not abandoned) the original vision for a common interchange format. EpiDoc now aims also to be a mechanism for the creation of complete digital epigraphic editions and corpora. We will argue that EpiDoc represents a better digital abstraction of the Leiden conventions than is achievable by a simple mapping Leiden's syntax for printed publication into digital form. A full EpiDoc document may contain, in addition to the text itself, information about the history of the inscription, a description of the text and its support, commentary, findspot and current locations, links to photographs, translations, etc. (Cayless et al. 2009).

As a result, they argue that the use of EpiDoc can support the creation of more sophisticated digital editions and digital corpora of inscriptions. In addition, the EpiDoc project has also created tools to convert Leiden-formatted inscriptions automatically into EpiDoc XML versions.

The Leiden conventions specify how inscription features besides the text should be represented in *print* and provide standard symbols that can be used to "convey the state of the original document and the editor's interpretation of that document" (Cayless et al. 2009). Directly mapping Leiden print syntax to a digital form, however, presented a number of issues that were covered in detail by the authors. Cayless et al. also noted that

digitally representing the typographic features of Leiden was only a first step, however, because epigraphic texts should also be "fully queryable and manipulable" in a digital environment:

> By the term "queryable", we do not simply mean that the text may be scanned for particular patterns of characters; we mean that features of the text indicated by Leiden should be able to be investigated also. So, for example, a corpus of inscriptions should be able to be queried for the full list of abbreviations used within it, or for the number of occurrences of a word in its full form, neither abbreviated nor supplemented. One can imagine many uses for a search engine able to do these kinds of queries on text (Cayless et al. 2009).

The ability to do such searches that "leverage the structures" embedded within Leiden, according to Cayless et al. (2009), first requires marked up inscription text that could then be parsed and converted into data structures that could be used to support the operations listed above. Such parsing requires lexical analysis that produces token streams that can then be fed into a parser, which can then produce parse trees that can be acted upon and queried in different ways. The authors granted that while EpiDoc is only one "possible serialization of the Leiden data structure" it does have the added advantage of having many tools available to already work with it.

Rather than making use of standards such as EpiDoc, Cayless et al. stated that the databases that supported most online epigraphy projects typically included various metadata fields and a large text field with the Leiden inscription directly transcribed without any markup or encoding (a fact supported by the survey in this review). The convenience of such a database setup is that it permits various fielded and full text searches, it is easy to connect with web-based front ends for forms, data can be easily extracted using Structured Query Language (SQL), and data can also be easily added to these systems. This makes it easy to insert new inscriptions as they are discovered. Nonetheless, this standard database approach has two major flaws according to Cayless et al.: 1) in terms of digital preservation, each "digital corpus" or database does not have distributed copies as a print corpus does; 2) these databases lack the ability to "customize queries" and thus "see how result sets are being constructed."

Another significant issue is that the way databases or their interfaces are designed can greatly influence the types of questions that can be asked. Making arguments similar to Dunn (2009) and Bodard and Garcés (2009), Cayless et al. argue that technical decisions such as the creation of a database are also "editorial and scholarly decisions" and that access to raw data is required in order to provide users the ability to both examine and correct decisions. Long-term digital repositories for inscriptions thus have at least two major requirements: the ability to export part or all of the data in standard formats and persistent identifiers (such as DOIs) at the level of a digital object so that they can be used to cite these objects independent of the location from where they were retrieved. As Cayless et al. explain, in a future where published digital inscriptions may be stored in various locations, the ability to cite items using persistent identifiers will be very important. Ultimately they see EpiDoc as a key component of such a future digital infrastructure for epigraphy, because it could serve not only as an interchange format, but also as a means of storing, distributing and preserving epigraphic data in a digital format.

All of these arguments, however, essentially lead the authors to one fundamental conclusion about epigraphy, that inscriptions are *texts* in complex environments, not just physical objects:

> This fact argues for treating them from the start as complex digital packages with their own deep structure, history, and associated data (such as images), rather than as simple elements in a standardized collection of data. Rather than engineering applications on top of a data structure that does not correspond well to the nature of the source material, we would do better to construct ways of closely representing the physical and intellectual aspects of the source in digital form, and then find ways to build upon that foundation (Cayless et al. 2009).

As indicated above, much of the earlier research that focused on inscriptions has investigated their archaeological context. The arguments made by Cayless et al. emphasize the need for inscriptions to be considered not just as simple data elements but also as complex digital objects with both a text inscription and an archaeological context. For epigraphy databases to support the growing field of digital epigraphy, Cayless et al. concluded that a mass of epigraphical data would need to be made available and that better tools would also be needed to gather, analyze and publish that data.

Despite Cayless et al.'s strong arguments in favor of greater adoption of EpiDoc, research by Bauer et al. (2008) has countered that EpiDoc has some specific limitations, particularly in regards to the development of

philological critical editions. According to the project website Hypereidoc [245] is an "XML based framework supporting distributed, multi-layered, version-controlled processing of epigraphical, papyrological or similar texts in a modern critical edition." The authors suggested that EpiDoc has limitations in terms of its expressive power and how individual results can be combined to form a cooperative product. They argue, therefore, that their proposed Hypereidoc framework provides an "XML schema definition for a set of annotation-based layers connected by an extensive reference system, validating and building tools, and an editor on-line visualizing the base text and the annotations" (Bauer et al. 2008). Their framework has been successfully tested by philologists working on the Hypereides palimpsest. [246]

The creation of digital transcriptions of epigraphic or papyrological texts, according to Bauer et al. requires a model that supports *multiple* levels of *annotation* to a base text:

> Annotations may mark missing, unreadable, ambiguous, or superfluous parts of text. They should also quote information about the *reason* of the scholar's decision e.g. other document sources, well-accepted historical facts or advances in technology. Annotations also provide meta-information about the author of the individual critical notes and expose the supposed meaning according to the given scholar. It is of a primary importance that no information should be lost during the transcription process (Bauer et al. 2008).

Although they noted that the Leiden conventions are the most accepted set of rules and that EpiDoc did successfully meet some of their needs, they also argued that digital critical editions would require a base text layer that always remained untouched. They developed a text model for annotation, VITAM (Virtual Text-Document Annotation Model) that contains "*virtual text-documents* as data items and *annotation sequences* and virtual text documents' *merging* as operations." Their multi-layered XML schema model is based on TEI and EpiDoc and defined a "base text layer" that stored just the original text and its physical structure, an "ordering and indexing layer" that defined page order and their place in codices, and one or more "Annotation layers" that store attached philological metadata. This model they argue better supports the creation of collaborative critical editions:

> Philologists can define their own Annotation Layers which may refer to only the Base Text Layer or one or more Annotation Layers. They can add notes and annotations to the original text and to previous annotations, they can make reflections on earlier work or create a new interpretation. We have designed a schema to handle these references and to support the distributed and collaborative work with using more Annotation Layers in one edition (Bauer et al. 2008).

At the same time, the authors noted that in order to make exact references to any point in the text, they needed to be able to describe the physical structure of the text. The "base text layer" was stored as a basic TEI document and as the palimpsest provided an existing physical structure with "codices, quires, leaves, sides, columns and lines" these were used as the primary structure for their reference system in order to define exact references to specific document parts. Such references were needed to support "philological processing," text annotation, and mapping between images and transcription. The authors do not discuss if they considered using CTS references.

Bauer et al. argued that TEI P4 and EpiDoc were less useful than the Hypereidoc model in the creation of philological annotations because they required that such annotations be stored in the form of XML tags inserted into a document, thus necessitating that annotations could only be embedded by philologists if the tags were balanced (due to the need for a well-formed XML document). Their proposed solution was to develop a reference system based on the *physical structure* of the document. "This enables the handling of any overlapping annotation," Bauer et al. stated, "With this reference system missing word and sentence boundaries can easily be described, even if interpreted differently by various philologists. Punctuations missing from the document can also easily be coded."

The unique nature of the palimpsest, however, with secondary text written over other original texts led them to define the text of Archimedes and Hypereides as the "undertext" and the new texts that were written over them as the "overtext." The page numbering of the "overtext" was used as the base for their reference system and they also defined their "ordering and indexing layer" independently from the "base text layer" and stored this data in an external XML file, noting that philologists would not necessarily agree on a page order and might

---

[245] http://hypereidoc.elte.hu/

[246] The Hypereides palimpsest is part of the larger Archimedes palimpsest, and for some of this work on Hypereides see (Tchernetska et al. 2007). The XML transcription of Hyperides can also be downloaded at http://hypereidoc.elte.hu/hypereides/downloads/hypereides-full.xml

want to use their own "ordering and indexing layer." While the "base text layer's" physical structure was based on the overtext, only pages were identified with overtext leaf and side while columns and lines were marked using the undertext so that the lines of Hypereides text could be specifically identified.

The Hypereidoc reference system supported three types of references: absolute references that point at a character position in the base text, internal relative references that point to a character position in the text "inserted by a previous annotation in the same annotation layer" and external relative references that point to a character position in the text "inserted by an annotation in a previous Annotation Layer." Several types of annotation are supported including embedded and overlapping annotations. They then developed a customized system that made use of the XML Pointer Language (XPointer), [247] which allows one "to point to an arbitrary point in an XML document." While TEI P5 has developed specifications for the use of XPointer, [248] Bauer et al. criticized these guidelines for only thinking about XPointer as a pointer to an *arbitrary tag* rather than an *arbitrary position* in a text, a feature that did not support the type of overlapping annotation that they needed. Nonetheless, they wanted to maintain maximum compatibility with TEI P5 so they made use of the <app> and <note> tags and publish an additional "flat file format" of their publications that does not make use of XPointer.

Bauer et al. also argued that another important feature supported by their system is effective version control. The base text-layer is read only and all annotation layers are modeled as separate sequences. In practice, they use a web server that handles all service requests in a RESTful manner. [249] The "virtual-text documents" are considered to be resources that can have the following version control operations, list (which shows the base text layer), create (which adds a new and time stamped annotation), and show (which gets the appropriate version of the file). To support the creation of digital critical editions, Bauer et al. have also designed their own XML editor [250] that manages their custom XML schema and supports working with both the layered XML file created using the Hypereidoc schema and the flat file format used to create compatible TEI P5 documents. In sum, Bauer et al. illustrated the importance of developing schemas and tools that support the representation of multiple scholarly arguments in the creation of digital critical editions of texts whether they are epigraphic or papyrological.

Online Epigraphy Databases

The sheer breadth of material available online and the active nature of this discipline is illustrated by the community maintained blog Current Epigraphy, [251] which reports news and events in Greek and Latin epigraphy and also publishes workshop and conference announcements, notices of new discoveries and publications and also provides descriptive links to digital epigraphy projects. Digital epigraphy projects are greatly varied, some include a small number of inscriptions from a particular area [252] while others include selected inscriptions in only Greek [253] or Latin [254] (typically with a chronological, geographic or thematic focus), and finally, some small projects focus on a particular type of inscription. [255] This subsection will provide an overview of a number of the larger projects [256] and relevant research that explores the major challenges facing this field in the digital world.

One of the largest Latin inscription projects available online is the Corpus Inscriptionum Latinarum (CIL), [257] a website that provides descriptive information regarding the CIL publication series and limited digital access to

[247] http://www.w3.org/TR/xptr-framework/
[248] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SATS
[249] REST stands for "REpresentational State Transfer" and RESTful web services have "a key design idiom that embraces a stateless client-server architecture in which the web services are viewed as resources and can be identified by their URLs. Web service clients that want to use these resources access a particular representation by transferring application content using a small globally defined set of remote methods that describe the action to be performed on the resource. REST is an analytical description of the existing web architecture, and thus the interplay between the style and the underlying HTTP protocol appears seamless." http://java.sun.com/developer/technicalArticles/WebServices/restful/
[250] http://hypereidoc.elte.hu/?i=editor/index
[251] http://www.currentepigraphy.org/
[252] Such as the Cyprus Inscriptions Database (http://paspserver.class.utexas.edu/cyprus/)
[253] See for example, "Poinikastas: Epigraphic Sources for Early Greek Writing", http://poinikastas.csad.ox.ac.uk/
[254] For example, see "Images Italicae" (http://icls.sas.ac.uk/imaginesit/)
[255] See for example, "Curse Tablets from Roman Britain", http://curses.csad.ox.ac.uk/index.shtml
[256] For a growing list of projects, see http://delicious.com/AlisonBabeu/clir-review+epigraphy
[257] http://cil.bbaw.de/cil_en/dateien/forschung.html

some of the inscriptions. Theodor Mommsen first formed the CIL in 1853, with the purpose of collecting and publishing all Latin inscriptions in an organized and scientific manner, and the publication of new and reissue of edited volumes still continues. [258] The CIL contains Latin inscriptions from the entire former Roman empire and publications were arranged by region and inscription type. Electronic access to a "collection of squeezes, photographs and bibliographical references maintained by the CIL research centre, sorted by inscription-number" is provided through the "Archivum Corporis Electronicum." [259] The database can only be searched by CIL volume and inscription number, but records for individual inscriptions can contain digital images of the inscriptions and squeezes as well as a selected bibliography. A variety of other resources are available from this website including a glossary of stone types used for inscriptions and a concordance to the CIL.

Another significant database of Latin inscriptions is the Epigraphik Datenbank Clauss-Slaby (EDCS) [260] that according to the website "records almost all Latin inscriptions." As of April 2010, the EDCS included over 539,000 sets of data for 381,170 inscriptions from over 900 publications covering more than 19,000 places with over 32,000 pictures. Inscription texts are typically presented without abbreviations and as completely as possible with only a few special characters to indicate missing texts. A full list of the publications included in this database is provided and the EDCS provides coverage of inscriptions in the CIL and many other major corpora. Users can search for inscriptions by text (of the Latin transcription), publication name, province or place of inscription (or a combination of these). A full list of relevant publications is given for an inscription search, and each inscription record contains abbreviated publication information, the province and place of the inscription, and a Latin transcription. Links are occasionally provided to images of these inscriptions in other databases.

The largest database of Greek inscriptions online appears to be the Packard Humanities Institute (PHI) Greek Inscriptions, [261] a project managed by the Greek Epigraphy Project at Cornell University with significant support from Ohio State University. The website describes the collection as a "scholarly work in progress" and is frequently updated. The Greek inscriptions are organized by 15 regions (Attica, Asia Minor, North Africa, etc.) and after selecting a region the user chooses from a variety of options such as main corpora, regional and site corpora, miscellaneous collections, miscellaneous books, journals, etc. By choosing one of these options, the user is then presented with a browsable list of inscription numbers, choosing one of these numbers then allows the user to view the entire Greek inscription. This whole collection or specific regions can also be searched in either Greek or Latin. In addition, there is also a concordance feature where a user can type a search pattern and the keyboard emulates an Ibycus keyboard and they can then launch a concordance search for that term.

One of the largest epigraphy resources for both Greek and Latin is EAGLE (Electronic Archive of Greek and Latin Epigraphy), [262] a federated database that searches across four epigraphical archives: Epigraphische Datenbank Heidelberg (EDH), [263] Epigraphic Database Roma (EDR), [264] Epigraphic Database Bari (EDB) [265] and Hispania Epigraphica (HE). [266] Each of these individual databases contains inscription texts, metadata and in some cases images for Greek and Latin inscriptions. While the EDB contains inscriptions from Rome only, EDR includes Latin inscriptions from Rome and greater Italy, and the HE contains Latin inscriptions and images from Spain. The EDH is a far larger database that contains both Latin and Greek inscriptions. EAGLE provides federated searching of these four databases and clicking on search results takes the user to the individual epigraphy databases. The website and searching are currently available only in Italian, but there are plans to create English, German, French and Spanish interfaces

[258] Currently there are 17 volumes in 70 parts (holding 180,000 inscriptions) with 13 supplementary volumes that include illustrations and indices, see http://cil.bbaw.de/cil_en/dateien/cil_baende.html. Websites have also been created for some individual volumes, such as a new edition of *Corpus Inscriptionum Latinarum*, vol. II: *Inscriptiones Hispaniae Latinae* (http://www2.uah.es/imagines_cilii/)
[259] http://cil.bbaw.de/cil_en/dateien/hilfsmittel.html
[260] http://www.manfredclauss.de/gb/index.html
[261] http://epigraphy.packhum.org/inscriptions/main
[262] http://www.eagle-eagle.it/
[263] http://www.uni-heidelberg.de/institute/sonst/adw/edh/index.html.en
[264] http://www.edr-edr.it/
[265] http://www.edb.uniba.it/
[266] http://www.eda-bea.es/pub/search_select.php?newlang=en

The EDH database that can also be accessed through EAGLE is a long-standing project in its own right that seeks to integrate Latin inscriptions from all parts of the Roman Empire into an extensive database. Since 2004, the EDH has also entered Greek inscriptions from this same chronological timespan. The EDH consists of three databases: the Epigraphic Text Database, the Epigraphic Bibliography (EBH) and the Photographic Database. While the Epigraphic Text Database contains over 56,000 inscriptions including many that were published outside of the standard major editions, the EBH contains 12,000 bibliographic records concerning monographs, journal articles and other sources of secondary literature regarding inscriptions in the EDH, and the Photographic Database includes over 11,000 photographs of inscriptions from various countries. While all three of these individual databases can be searched separately, photos and bibliographic information are often found within inscription records. The records for inscriptions in the EDH are very detailed and typically include a unique EDH identifier, images, historical and physical information as well as transcriptions. Currently the EDH is also participating in two major data integration projects, the previously mentioned EAGLE and the Concordia Initiative.

ConcEyst [267] (short for Das Eichstätter Konkordanzprogramm zur griechischen und lateinischen Epigraphik) is another large database of Greek and Latin inscriptions (from the Roman province of Pontus-Bithynia) available for download online. This database is continuously updated and both it and a search interface in concordance format can be downloaded.

While a large number of epigraphic resources online are dedicated to Greek and Latin, there are also major epigraphic resources for inscriptions in other languages. The Bibliotheca Alexandrina has created a "Digital Library of Inscriptions and Calligraphies" [268] that is in the process of creating a comprehensive digital collection of inscriptions in Ancient Egyptian (Hieroglyphic, Hieratic, Demotic and Coptic scripts), Arabic, Turkish, Persian and Greek. This collection also includes inscriptions bearing the Thamodic, Musnad, and Nabatean scripts. This whole collection of inscriptions can be searched and can also be browsed by language. For each inscription record, a digital image of the object, building or monument from which the inscription came is provided along with a description and historical information regarding the object, the text of the inscription on the object, a transliteration, and often an English translation. This website's presentation of both the archaeological object on which the inscription is found along with a full text transliteration helps to emphasize the unique nature of inscriptions as both texts and archaeological objects with a context.

Another major epigraphical project with a significant concentration outside of Greek and Latin is the "Inscriptions of Israel/Palestine," [269] which seeks to "collect and make accessible over the Web all of the previously published inscriptions (and their English translations) of Israel/Palestine from the Persian period through the Islamic conquest (ca. 500 BCE - 640 CE)." According to the website, there are about 10,000 such inscriptions, written primarily in Hebrew, Aramaic, Greek and Latin, by Jews, Christians, and pagans. These inscriptions are quite varied and have never been collected or published in a systematic fashion. This project is a collaborative effort entirely supported by Brown University and a variety of other partners, and their ultimate goal is to gather all of these inscriptions together and publish them online as a scholarly resource. Each inscription will be converted into a tagged XML text that is compatible with EpiDoc, but this project's markup and DTD have some differences:

> Rather than treating the text as the primary object (with the goal of moving it relatively easily to publication), our mark-up treats the inscribed object as primary. The metadata, which is specified with great detail to allow for database-like searching, is all put into the Header. The Header contains information such as type of object; date range; locations (present, find, and original); type of inscription; language, etc.). Individual "div" sections contain the diplomatic and edited version of the texts, in their original languages, and an English translation. The source of each is always acknowledged. The DTD already contains a scheme for richly marking-up the contents of the texts themselves. We have recently decided only to mark textual and editorial features; content such as names, places, occupations, etc., will be added (perhaps in part through automated processes) at a later stage. Almost all of our tags follow, to the extent possible, the accepted TEI and EpiDoc usages.

[267] http://www.ku-eichstaett.de/Fakultaeten/GGF/fachgebiete/Geschichte/Alte%20Geschichte/Projekte/conceyst/
[268] http://www.bibalex.org/calligraphycenter/InscriptionsLibrary/Presentation/index.aspx?Lang=en
[269] http://www.stg.brown.edu/projects/Inscriptions/

As the project creators also comment, they would like to support a higher level of encoding but limited staffing and funding options have made this impossible. A database of about 1000 inscriptions is currently available, and users can search the inscription metadata, the text of the inscription and English translation, or both at the same time using a variety of options.

Finally, another major resource is Inscriptifact, [270] an image database of inscriptions and artifacts that has been created as part of the West Semitic Research Project at the University of California. This database provides access to high-resolution images of inscriptions (papyri, incised inscriptions on stone and clay, cuneiform tablets, stamp seals, etc.) from the both the Near Eastern and Mediterranean Worlds. Currently the archive contains over 250,000 images and is updated regularly. For access to the database, an application form must be faxed. A recent article by Hunt et al. (2005) has also described the creation of this database and the standards used in detail. InscriptiFact made use of many historical photographs that were often the only source of information for many inscriptions, but also utilized a number of advanced photographic techniques to better capture images of inscriptions on all types of objects and monuments.

One major challenge was that fragments of different inscriptions or collections of fragments were often scattered among various museums, libraries and archaeological collections. [271] In order to address this issue, Hunt et al. advised that a fairly complicated cataloging and data model needed to be developed:

> Data in InscriptiFact are organized around the concept of a text, rather than a digital object or a collection containing texts. A "text" in this context is a virtual object in that a given text may not physically exist at any one place in its entirety. That is, since text fragments are often found in scattered locations in various collections, InscriptiFact brings together images of a given text regardless of the location of individual parts of that text in institutions around the world (Hunt et al. 2005).

The authors argued that both FRBR and unqualified Dublin Core [272] did not readily represent the type of metadata required by scholars of ancient texts. They noted that for such scholars "a text is an intellectual concept" and scholarly cataloging must be created for *all* of its manifestations and include information about the physical objects that contain the inscription, the "intellectual work of the inscription itself," as well as photographic images and digital images. Hunt et al. suggested that the FRBR concept of the work could not be utilized for ancient inscriptions since scholars only have the inscribed physical object (manifestation) and that given texts may have been subdivided into many physical fragments. "It is the job of archaeologists, linguists, epigraphists, philologists, and other specialists to try to reconstruct the original text," Hunt et al. (2005) maintained, "that is, figure out what pieces fit together, how the text is organized, when it was inscribed, and what, in fact, the intellectual content of the inscription might be." They thus decided that it was not useful to separate the intellectual work of the text from the physical object or objects upon which it was inscribed.

Within InscriptiFact, the intellectual work has been defined as the inscription within the *context* of the physical object where it was inscribed. Since inscriptions do not have expressions as defined by FRBR, they used the Dublin Core element *relation* to map relationships between the textual content of an inscription and instances of that archaeological/physical context. While the basic objects that are delivered to users are digital images and would seem to correspond to FRBR items, Hunt et al. insisted that images of complicated objects (e.g. a plate of fragments) that can include multiple texts illustrates that with inscriptions there can be a "many to many" relationship between works and items. The final approach that they adopted was to separate cataloging for the text (the inscription or work) from the images (the digital objects or items) and they extended qualified Dublin Core to "include an additional qualifier to denote manifestation."

In addition to federated databases and digital collections of inscriptions, there are also a number of reference tools now available online that have been created to assist scholars in finding inscriptions. The CLAROS: Concordance of Greek Inscriptions [273] database provides access to a computerized concordance of editions of ancient Greek inscriptions that have been published in the last 100 years. The fifth edition which was last

---

[270] http://www.inscriptifact.com/

[271] This is similar to the problem communicated by Ebeling (2007) regarding the need to create "composite" texts for Sumerian cuneiform tablets that were physically fragmented

[272] http://dublincore.org/

[273] http://www.dge.filol.csic.es/claros/cnc/2cnc.htm

updated in 2006 includes more than "450,000 equivalences coming from more than 750 collections." Currently this concordance provides limited links between the results of bibliographic searches with electronic versions of inscriptions in the Inscriptions of Aphrodisias, the Greek Epigraphy Project of the PHI, and some texts from Egypt published in the Duke Data Bank of Documentary Papyri. A full list of collections that are included in the concordance is provided as well as a useful list of abbreviations used for classical and epigraphical publications. The sheer breadth of this database illustrates how many Greek inscriptions have been published in multiple editions, and the consequent challenges of integrating links to electronic versions of these inscriptions in other databases.

EpiDoc-Based Digital Epigraphy Projects

The Inscriptions of Aphrodisias (ALA 2004) [274] website provides access to the electronic second edition of "Aphrodisias in Late Antiquity: The Late Roman and Byzantine Inscriptions" by Charlotte Roueché of King's College London. This website provides access to a second edition that has been expanded and revised from the version published by the Society for the Promotion of Roman Studies in 1989. Charlotte Roueché (2009) has explained the process of creating this website in detail and as noted by Cayless et al. (2009) earlier, also highlighted the point that inscriptions have two identities, both as a text and as an "archaeological object with a context." Despite this identity as a text, Roueché remarked that inscribed texts had often been omitted from the literary canon. As an example, she imparted that there were two verse inscriptions from Aphrodisias on the same block, but only one was quoted in the *Greek Anthology* and thus ended up in the TLG, while the other never entered the literary tradition.

While ALA 2004 includes about 2000 inscriptions, Roueché stressed that the sheer scale of inscriptions almost *necessitates* electronic publication. She also reported that she was first introduced to EpiDoc by her colleagues (Tom Eliott and Charles Crowther) and thus felt that one important question to consider was how to bring together domain specialists with the technical experts that can help them. In order to begin this process, Roueché and others started the Epidoc Aphrodisias Project (EPAPP) [275] in 2002 to develop tools for presenting Greek and Latin inscriptions on the Internet using EpiDoc. The project held two workshops in the United States and United Kingdom to get input from interested experts, and the initial outcome of this project was ALA 2004. After securing more grant funding, the project expanded ALA 2004 and also published IAph 2007 and the whole corpus is referenced as InsAph. As ALA 2004 is considered to be the second edition of her book, Roueché articulated that the website that has been produced is stable and all inscriptions are citable. Furthermore, Roueché believed that simply creating ALA 2004 was important in order to demonstrate what was possible for an electronic publication of inscriptions. Nonetheless although she obtained an ISBN for the website, she had trouble getting librarians at her institution to create a catalog record for it, thus reinforcing the idea that the website was not a real publication, a problem also reported by archaeologists interviewed by Harley et al. (2010).

Despite this difficulty, even more daunting were the challenges of data integration between different epigraphy projects. As demonstrated through even the brief survey conducted by this review, there are numerous epigraphy projects online, and Roueché reported that more pioneering work in digital epigraphy has involved Latin inscriptions. Nonetheless, as listed above, one of the major Greek inscriptions projects is PHI Greek Inscriptions, which also contains all the inscriptions from Aphrodisias published through 1993. In the future, Roueché is hoping to embed PHI Greek identification numbers in the XML of inscriptions in ALA 2004 so the PHI website can automatically receive updated information from ALA 2004 if it changes.

Fundamental to the problem of data integration, Roueché asserted, is convincing more epigraphists to take up the EpiDoc standard. One means of doing this she concluded was to demonstrate how EpiDoc was not a radical shift, but rather an extension of how epigraphists have always worked:

[274] http://insaph.kcl.ac.uk/ala2004/index.html
[275] http://www.epapp.kcl.ac.uk/

> The aim is to get epigraphers to perceive that EpiDoc encoding simply represents an extension of the approach which produced the Leiden conventions. As often in humanities computing, it is important to demonstrate that the intellectual activities and processes in what appear to be separate fields are in fact closely related (Roueché 2009, pg. 165).

Another problem Roueché admitted was that many humanities scholars simply wanted to define a problem and then have technicians solve all of the challenges in creating a digital solution, a situation she rightly concluded was simply not viable. Key to solving this difficulty is the development of a common language, for she noted that both epigraphists and computer scientists have their own acronyms (CIL, XML). The most critical task, however, Roueché insisted, is to demonstrate the added scholarly value of electronic publication to epigraphists. Among the many benefits of electronic publication, perhaps the most significant Roueché listed was the ability of electronic publication to better accommodate the interdisciplinary nature of inscriptions as both literary texts and archaeological objects. The new ability to both disseminate and integrate inscriptions not only with other collections of inscriptions but also with papyri, manuscripts, or seals, Roueché hoped would help "break down what have been essentially false distinctions between texts which all originate from the same cultural milieu, but are recorded on different media" (Roueché 2009, pg. 167). As illustrated throughout the earlier discussion of archaeology, the ability to reintegrate the textual and material records in a digital environment is a critical issue that must be addressed. In addition, Roueché also suggested that is far easier to update a digital corpus of inscriptions and for scholars to work collaboratively. These new forms of collaboration, however, also lead to questions regarding data ownership and authorship credit. Indeed, many of the greatest challenges may be social rather than technical Roueché concluded.

The second major part of the InsAph corpus is IAph 2007, [276] which provides access to the first edition of an online corpus of inscriptions from Aphrodisias that were recorded up to 1994. All of the editions, translations and commentary have been provided by Joyce Reynolds, Charlotte Roueché and Gabriel Bodard. In addition, the inscriptions have been marked up using EpiDoc and individual XML files for inscriptions or the entire XML repository of inscriptions can be downloaded from the site along with a DTD. [277] Slightly more than 1,500 inscriptions (a 1/3 of which have not been previously published) are available through the database, and the whole collection can either be searched by free text (Greek, Latin and English in transcriptions and editions) or by category (such as date, object text and text type). Browsing access to the inscriptions is also provided through different "Tables of Contents" including locations, date, text categories, monument types, decorative features and texts new to the edition. Interestingly, this database also provides a number of indices to the inscriptions including Greek words, Latin words, personal and place names, and several other characters and features. There is also a "Plan of Aphrodisias" [278] that allows users to choose a section of the city that then provides them with a list of inscriptions in that area. Each inscription record includes extensive information, multiple images, the Greek or Latin text, a diplomatic transcription, an EpiDoc XML file, and an English translation. In addition, each inscription has a citable and permanent URL. [279]

Two recent articles by Gabriel Bodard (Bodard 2008, Bodard 2006) have looked at some of the issues regarding the benefits and difficulties of creating electronic publications of inscriptions such as IAph 2007. Bodard listed six features for analysis in terms of the opportunities of electronic publication: accessibility, scale, media, hypertext, updates, and iterative research and transparency. The digital publications of the inscriptions of Aphrodisias were the first major ones to adopt EpiDoc, Bodard explained, and the use of this standard based on a subset of the TEI guaranteed maximum compatibility with many other digital humanities projects. One important thing to note, Bodard declared, was that:

> An essential concept behind EpiDoc is the understanding that this form of semantic markup is not meant to replace traditional epigraphic transcription based on the Leiden conventions. The XML may (and almost inevitably will) encode more information than the range of brackets and sigla used in Leiden, but there will always be a one-to-one equivalence between Leiden codes and markup features in the EpiDoc guidelines (Bodard 2008).

[276] http://insaph.kcl.ac.uk/iaph2007/index.html
[277] http://insaph.kcl.ac.uk/iaph2007/inscriptions/xml-repo.html
[278] http://insaph.kcl.ac.uk/iaph2007/reference/plan.html
[279] http://insaph.kcl.ac.uk/iaph2007/iAph040113.html#edition

As was submitted previously by both Cayless et al. (2008) and Roueché (2009), Bodard also argued that encoding inscriptions in EpiDoc is a natural extension to the work that epigraphists already do.

To return to the six features listed by Bodard, he argued that the most obvious benefit of electronic publication was accessibility, and that publishing inscriptions online both serves as scholarly outreach and fosters interdisciplinarity. Secondly, as was also argued by Roueché, Bodard advocated that scale was one of the most significant opportunities for electronic publishing. He revealed that unlike with the first printed edition, ALA 2004 was able to include multiple photographs of each inscription, thus helping to better set the inscriptions in their archaeological context. The scale of digital publishing also allowed for the inclusion of far more interpretative material and for the explanation and expansion of text that had once needed to be abbreviated by epigraphists and papyrologists for print publication:

> …. once the restrictions of a page limit are removed this abbreviated text can be expanded, conventions can be glossed, descriptions of comments can be repeated where they are relevant, and cross-references can be made more self-explanatory. This is not necessarily to reject generations of scholarship and academic jargon, which is familiar to practitioners of our disciplines and serves a useful function of communication in addition to space-saving. Rather, by expanding, explaining, and illustrating these conventional sigla and abbreviations we are enhancing our scholarly publication by making it more accessible to outsider (Bodard 2008).

Making inscriptions more accessible to a general audience is perhaps one of the greatest benefits to publishing inscriptions online.

The third feature listed by Bodard was media, and in addition to the far larger number of photographs, Bodard also suggested that digital reconstruction of monuments with inscriptions could prove very useful. In addition, increased levels of geographical access can now be provided through digitizing maps and plans that can then be hyperlinked to editions of inscriptions. One potential avenue for exploration Bodard outlined would be the use of a global mapping API [280] such as that provided by Google Maps [281] "to plot not only findspots and ancient and modern locations of finds, but also places and other geographical entities named or implied in the texts themselves" (Bodard 2008). Both Eliott and Gillies (2009a) and Jeffrey et al. (2009a) have also suggested the utility of using modern mapping technology such as GIS and historical named entity recognition techniques to provide better access to historical materials in archaeology and classical geography.

Perhaps the most significant impact of electronic publication on inscriptions, however, Bodard argues, is through his fourth feature, hypertext, with its myriad possibilities of supporting sophisticated linking and dynamic referencing. Internal hyperlinks within a publication, as Bodard noted, enabled making stronger links between data, narrative commentary and other supporting materials. New ways of navigating the material become available as a user can go from a narrative commentary directly to an inscription, or from an inscription of interest directly to commentary. External hyperlinking to other projects (such as other inscription collections and secondary reference works) also offers powerful possibilities. Another important aspect of hyperlinking Bodard underscored was the potential of dynamic linking or "live hypertext sharing." As both InsAph 2007 and ALA 2004 provide downloadable EpiDoc XML files both for the individual inscriptions and the entire corpus and also provide "transparent and predictable URLs for dynamic linking" (Bodard 2008), other projects can both easily link to individual inscriptions or download the entire corpus for reuse. As far as Bodard knew, however, no other projects had made reuse of any of the EpiDoc files available as yet.

The fifth feature listed by Bodard, that of the ability to easily update data was also briefly discussed by Roueché. Although the possibility of instantaneous updating can be very useful, Bodard also proposed that some electronic publications may need to be kept "static" not only due to the burden on the author but also because of the need for a stable and citable publication "that has to interact with, and be reviewed within, the world of peer reviewed, cited, traceable, and replicable scholarship." Consequently all versions (with their URLs) of an inscription must be maintained even if changes or corrections are made, in case a scholar has cited an earlier

[280] An API, short for "Application Programming Interface" is a "set of routines, protocols, and tools for building software applications" (http://webopedia.com/TERM/A/API.html). According to Webopedia, a "good API makes it easier to develop a program by providing all the building blocks." Programmers can then use these "building blocks" to more easily write applications that are consistent with a particular operating environment or software program.

[281] http://code.google.com/apis/maps/index.html

version of an inscription. For these reasons as well as the challenges of project-based funding, the authors decided to create ALA 2004 and IAph 2007 not as "living databases" but as more traditional "one-off publications."

The final feature analyzed by Bodard, that of the ability of electronic publication to support iterative research and transparency, was also previously discussed by this author and Garcés (2009) in terms of digital editions. The availability of source files and code makes research more replicable by other scholars for it provides access to primary source data, allows other scholars to examine the digital processes, markup and techniques used to create the collection, and also allows them to use their own algorithms and tools to create new digital editions. Such transparency is essential to all scholarly research, and Bodard concludes that:

> Even more central to the research process, however, is the fact that a true digital project is not merely the result of traditional classical research that is at the last minute converted to electronic form and made available online. Rather the XML files (in the case of *Inscriptions of Aphrodisias* and other EpiDoc projects, other data models for other types of project) that lie behind the publication, are the direct result of, and primary tools for, the academic research itself. These files contain the marked-up data, Greek or Latin texts, descriptions, editorial commentary and argumentation, references and metadata, all in machine-readable and actionable form. It is this single, structured collection of source data which is taken by the machine process and run through a series of XSLT stylesheets which generate, in turn, the web presentations of individual or groups of inscriptions, the contextual tables of contents, indices, concordances, prosopographical and onomastic tables, and so forth (Bodard 2008).

Thus the electronic files and source code that have been created for this publication are every bit as important if not more important than the final website created from them.

One closely related project to both ALA 2004 and IAph 2007 is the Inscriptions of Roman Tripolitania (IRT 2009). [282] IRT 2009 is the enhanced electronic reissue of a publication that first appeared in 1952 that has been created by Gabriel Bodard and Charlotte Roueché and is hosted by King's College London. Electronic publication has allowed for the inclusion of the full photographic record of the original print publication and the linking of the inscriptions to maps and gazetteers. As with IAph 2007, individual EpiDoc XML files and an entire XML repository of inscriptions can be downloaded. [283] There are a variety of ways to access the inscriptions. The chapters of the print publication have been made available online as text files with hyperlinks to the relevant inscriptions. In addition as with IAph 2007, various tables of contents (locations, dates, text categories, monument types) and indices (Latin words, Greek words, fragments of text, personal names, and other features) provide alternative means of access to the inscriptions. The website notes that indices to this edition were generated from the texts themselves and thus differ from the printed edition. Each inscription record is similar to those in IAph 2007 with one key difference, many of the records in IRT have "findspots" that have been linked to maps. [284] A map can be used to browse the collection of inscriptions with links provided to individual inscription records. Each inscription also has a citable and permanent URL.

Another related project that has recently begun is the Inscriptions of Roman Cyrenaica (IRCyr). [285] This website provides access to inscriptions gathered by Joyce Reynolds of Newnham College Cambridge between 1948 and the present. This project draws off the experience gained in publishing ALA2004 and IAph 2007 and they plan to both present the documents online in a similar fashion to these websites and to link all inscriptions to an online map of Roman Cyrenaica that is being prepared by the Pleiades project. No inscriptions database is currently available at this website. IAph 2007, IRT 2009 and IRCyr are all also participating in the Concordia project.

Finally, another project that makes partial use of EpiDoc is the U.S. Epigraphy project that is dedicated to collecting and digitizing Greek and Latin inscriptions, but in this case is focused on those preserved in the United States of America. [286] The project was founded at Rutgers University in 1995 and has been based at Brown University since 2003, where the present website was developed with help from the Scholarly Technology Group. Every inscription that has been catalogued by this project has been assigned a unique

[282] http://irt.kcl.ac.uk/irt2009/index.html
[283] http://irt.kcl.ac.uk/irt2009/inscr/xmlrepo.html
[284] For example, the map for the findspot "Sirtica", http://irt.kcl.ac.uk/irt2009/maps/index.html?ll=31.2,16.5&z=10
[285] http://ircyr.kcl.ac.uk/
[286] http://usepigraphy.brown.edu/

identifier or U.S. epigraphy number. The database of almost 2,500 Greek and Latin inscriptions can be browsed by publication or collection and searched for by language, place of origin, date, type of inscription, type of object, and material (among many other metadata categories) as well as by bibliographic information. According to the website, a "growing digital edition of the collection currently registers some 400 transcriptions of Latin texts encoded according to EpiDoc conventions and provides some 1,000 photographs and images of the inscriptions in our corpus." This makes the U.S. Epigraphy project one of the first major projects to begin the encoding of its texts in EpiDoc.

The Challenges of Linking Digital Epigraphy and Digital Classics Projects

As the above overview of inscription projects demonstrated, there are records of many of the same inscriptions in various databases, and many databases have used their own technological implementations to provide access to collections online. The sheer scale of many such projects and the growing number of inscriptions available online require computational solutions.

Recently Leif Isaksen has proposed the development of an "augmented reality mobile application" (such as for the IPhone) to support the "crowdsourcing" of epigraphy (Isaksen 2009). In theory, such an application could allow tourists or archaeologists to submit spatially-located images of inscriptions to a central inscription database that could also include a website where corrections and translations of inscriptions could be proposed based on multiple images of inscriptions. Creating such a central database of images would also support research work in advanced imaging techniques for various cultural heritage projects.

While some websites use EpiDoc, the challenges of linking between varying epigraphy databases and other digital classics resources such as papyrological databases is a growing challenge for which various solutions have been explored by projects such as the recently completed LaQuAT (Linking and Querying of Ancient Texts). [287] LaQuAT was a collaboration between the Centre for e-Research Kings College, London [288] and EPCC [289] at the University of Edinburgh. Two recent articles by Bodard et al. (2009) and Jackson et al. (2009) have described the basic goals and technological challenges faced by this project. The LaQuAT project used OGSA-DAI, [290] an open source distributed data management software, to successfully create a demonstrator that provided uniform access to different epigraphic and papyrological resources. Basically the LaQuAT project sought to build a proof of concept that explored the possibilities of "creating virtual data centres for the coordinated sharing of such resources" and examined how "distributed data resources can be meaningfully federated and queried."

From a preliminary analysis of digital classics resources, Jackson et al. reasoned that a data integration project would need to deal with the various complexities of annotated corpora, material in relational databases and large numbers of XML files. Such research is of growing importance due to the large number of individual and isolated digital resources that have been created. "In the fields of archaeology and classics alone," Bodard et al. (2009) explained, "there are numerous datasets, often small and isolated, that would be of great utility if the information they contained could be integrated." The researchers found that four major issues needed to be addressed in terms of potential integration: 1) the formats of resources were very diverse; 2) resources were often not very accessible (e.g. stored on individual department or scholar's computers), and even data published on websites was typically not available for reuse; 3) resources were available to be used only in isolation (e.g. single inscription databases); 4) resources were owned by different individuals and communities with varying rights schemes. The LaQuAT project thus wanted to explore if bridges could be built between different data silos in order to support federated searching at the least and they thus brought together experts in distributed data management and digital humanities.

[287] http://www.kcl.ac.uk/iss/cerch/projects/completed/laquat.html
[288] http://www.kcl.ac.uk/iss/cerch
[289] http://www.epcc.ed.ac.uk/
[290] http://www.ogsadai.org.uk/

The original plan of LaQuAT was to link three projects, Project Volterra [291] (an Access database of Roman legal texts and metadata), Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV) [292] (a "database of papyrological metadata in relational and TEI-XML format" that includes information on 55,000 papyri and is stored in FileMaker Pro), and the Inscriptions of Aphrodisias (IAph). These collections span about 500 years of the Roman Empire and also overlap in terms of places and people. While all of the datasets are freely available and both the IAph and HGV collections have been published as EpiDoc XML that can be downloaded under a CC Attribution License, it was the master databases of both the HGV and Volterra that were needed for this project and they had to be specifically requested (Bodard et al. 2009). Despite the initial desire to support cross database searching of all three projects, they found, however, that the challenges of integrating the relational databases were so complicated that they focused on simply Volterra and HGV in this project. One question they still wished to explore was if information in HGV could be used to reduce uncertainty regarding dates in the legal texts in Volterra. In order to integrate HGV and Volterra, they created annotations databases for each project or "randomly-generated values associated with each record in the original databases" so they could "demonstrate cross-database joins and third-party annotations" (Jackson et al. 2009).

The project used OGSA-DAI [293] for data integration because it was considered a de-facto standard by many other e-science projects for integrating heterogeneous databases, it was open-source and it was also compliant with many relational databases, XML and other file-based resources. OGSA-DAI also supported the exposure of data resources on to grids (Bodard et al. 2009). Most importantly, in terms of data integration:

> …OGSA-DAI can abstract the underlying databases using SQL views and provide an integrated interface onto them using distributed querying. This fulfils the essential requirement of the project to leave the underlying data resources untouched as far as possible (Jackson et al. 2009).

One essential goal of LaQuAT was to be able to support federated searching of a "virtual database" in order that the underlying databases would not have to undergo major changes for inclusion in such a resource. "The ability to link up such diverse data resources, in a way that respects the original data resources and the communities responsible for them," Bodard et al. 2009 asserted, "is a pressing need among humanities researchers."

A number of major issues complicated data integration, however, including data consistency and some specific features of OGSA-DAI. To begin with, some of the original data in the HGV database had been "contaminated by control characters," a factor that had serious implications for the OGSA-DAI system since it provided access to databases via web services, which are based on the exchange of XML documents. Since the use of control characters within an XML document results in an invalid XML file that cannot be parsed, they had to extend the system's "relational data to XML conversion classes to filter out such control characters and replace these with spaces." The Volterra database also presented its own unique challenges, particularly in terms of database design, since it was discovered that not all tables had the same columns and some columns with the same information had different names. A second major challenge was the lack of suitable database drivers, and the data from both Volterra and HGV were ported into MySQL to be able to interact with OGSA-DAI. Other issues included needing to adapt the way the OGSA-DAI exposed metadata and having to alter the way the system used SQL views because of the large nature of the HGV database. In the end, the project could only use a subset of the HGV database to ensure that query time would be reasonable. Despite these and other challenges the project was able to develop a demonstrator that provided integrated access to both HGV and Volterra. [294]

The LaQuAT project had originally assumed that one of the most useful outcomes of integrating the two databases would be where data overlapped (such as in terms of personal and place names), but they found instead that clear cut overlaps were fairly easy to identify. A far more interesting question they proposed instead

---

[291] http://www.ucl.ac.uk/history2/volterra/

[292] HGV is also federated as part of Trismegistos, http://aquila.papy.uni-heidelberg.de/gvzFM.html

[293] While the technical details of this software are beyond the scope of this paper, Jackson et al. explain that: "OGSA-DAI executes workflows which can be viewed as scripts which specify what data is to be accessed and what is to be done to it. Workflows consist of activities, which are well-defined functional units which perform some data-related operation e.g. query a database, transform data to XML, deliver data via FTP. A client submits a workflow to an OGSA-DAI server via an OGSA-DAI web service. The server parses, compiles and executes the workflow."

[294] For more on the infrastructure proof of concept design please refer to (Jackson et al. 2009). This demonstrator can be viewed at http://domain001.vidar.ngs.manchester.ac.uk:8080/laquat/laquatDemo.jsp

was to try and automatically recognize "the co-existence of homonymous persons or names in texts dated to within some small number of years of one another, for example" (Jackson et al. 2009). Historical named entity disambiguation thus presented both a major opportunity and challenge to data integration. In addition, another significant barrier to querying multiple databases was the problem of semantic ambiguity:

> To run queries across multiple databases, a researcher would already need a significant degree of understanding about what each database contained and also which tables and columns contained data that was semantically equivalent and could therefore be compared or tested for equality. Any such infrastructure would have to provide a far greater degree of support for making the databases seem as if they are indeed part of one virtual database, for example by normalizing dates (Jackson et al. 2009).

In addition to semantic ambiguity in terms of how data was described or stored, Jackson et al. also pointed out that once you started trying to automatically link humanities databases the fuzzy and interpretative nature of much of this data became quite problematic. Other more specific challenges included knowing when to join columns, variant names for historical entities, various ways of representing dates, the precision and uncertainty of dates, and errors in databases that cannot easily be changed.

One major conclusion reached by the LaQuAT project was that more virtual data centers needed to be created that could integrate several data sources and they were for this reason actively participating in the DARIAH project, hoping that the solutions LaQuAT had developed would

> ….have a lifespan beyond the initial project and will provide a framework into which other researchers will be able to attach resources of interest, thus building up a critical mass of related material whose utility as a research tool will be significantly greater than that of the sum of its parts. We see this project as providing an opportunity to start building a more extensive e-infrastructure for advanced research in the (digital) humanities (Bodard et al. 2009).

As part of this work, they hoped to convince scholars in different countries to abandon a data-silo mentality and help build up a large mass of open material. In terms of future research, they argued that far more research was needed into the issue of cross-database linking in the humanities, especially in the linking of relational and XML databases, which their project was unable to investigate further.

The scholarly importance of linking the study of inscriptions to other sources of archaeological or other material, particularly to help provide a greater context for individual inscriptions, has also been made by Charlotte Tupman (Tupman 2010). In her discussion of funerary inscriptions found on monuments, Tupman noted that different categories of funerary evidence (e.g. pottery, bone fragments, etc.) typically need to be assembled for fuller understanding of an inscription [295] and that there is no easy way to present the varied archaeological evidence, the funerary text and images of the monument it was found on in a way that is easily comprehensible to scholars. As funerary texts were rarely published with other related material evidence, Tupman observed that typically these inscriptions have not been thought of as archaeological material but as "the preserve of historians and literary scholars" since they are considered as "texts rather than artifacts" a point also made previously (Roueché 2009, Bodard et al. 2009).

Tupman argued that it would be highly desirable not just to link funerary inscriptions to images of the monument on which they were found, but to then link these monuments to *other* objects found in the same archaeological context. While Tupman granted that it certainly made some sense that inscriptions, pottery catalogues and bone analysis are published separately (as they are separate disciplines), she also contended that all data that was published should be able to be linked to at a minimum, and ideally, to be combined with other data:

> Specialists, therefore need to work to make their material available to others in a way that permits their various forms of data to be combined meaningfully. This will be most effective if undertaken collaboratively, so that shared aims and standards can be established. This does not imply that there should be any diminution of expert knowledge or information in any of these fields for the sake of making it easier for others to digest; to do so would entirely miss the point of the exercise. Rather, we should be seeking ways of linking these different types of information in a rational and useful manner that not only increases our own understanding of the data, but also enhances the way in which computers can process that data (Tupman 2010, pg. 77).

[295] Kris Lockyear has also made similar arguments about the importance of integrating numismatic evidence with other archaeological evidence.

Tupman thus encouraged both the use and creation of collaborative standards and to make data both human readable and machine actionable. While she suggested that perhaps Semantic Web technologies might be useful in this regard, Tupman also proposed that the use of XML and in particular EpiDoc to support the digital publishing of inscription data would be highly beneficial towards achieving these aims. After reiterating Gabriel Bodard's (Bodard 2008) six transformational qualities of digital publishing, Tupman listed a number of advantages of using XML including: incorporating marked up texts into databases, interlinking marked up inscriptions with other types of XML files, the ability of researchers to add their own markup, and the possibility of using XSLT to produce different displays of the same source file (e.g. for a beginning student vs. an advanced scholar). Tupman concluded that providing inscription data as EpiDoc XML files not only lessened editorial ambiguity (e.g. by supporting the encoding of variant readings) but could also allow researchers to use digital tools to sort large amounts of data and thus ask their own questions of the materials

Advanced Imaging Technologies for Epigraphy

While the above sections examined some of the difficulties in providing sophisticated access to inscriptions as digitized texts and of linking between collections, other research has focused on the challenges of advanced imaging for inscriptions as archaeological objects. This section will briefly examine several state-of-the-art approaches. [296]

The eSAD project has recently developed a number of image-processing algorithms for studying ancient documents that have also been made available to scholars through a portal (Tarte et al. 2009). Using images of wooden stilus tablets from Vindolanda as their testbed, they developed algorithms that helped to rebalance illumination and remove wood grain. This project then extended the data model and interface of the previously developed Virtual Research Environment for the Study of Documents and Manuscripts (VRE-SDM) [297] so that it could call upon these algorithms using web services that make use of the UK National Grid Service. This work served as a proof of concept for the viability of the VRE-SDM and supported the development of a portal that hid the technology from scholars, utilized the grid and web services as a means of providing access to powerful image processing algorithms, and also allowed the eSAD to disseminate its results to both classicists and image processing researchers. The VRE-SDM also supported scholars that wanted to collaborate by providing them with a virtual environment where work could be shared.

In addition to image processing algorithms for digital images of inscriptions, other research has developed advanced 3D techniques to capture better images of squeezes taken of inscriptions. Barmpoutis et al. (2009) have asserted that conventional analysis of ancient inscriptions has typically been based on observation and manual analysis by epigraphists, who both examine the lettering and attempt to classify inscriptions geographically and chronologically. One particular method that has been traditionally used is where researchers "use a special type of moisturized paper (squeeze) which they push on the inscribed surface using a brush specially adapted for the purpose. When the letters are shaped on the squeezed paper, the archaeologists let it dry, creating that way an impression of the inscription" (Barmpoutis et al. 2009). The authors reported that many collections of squeezes have been created [298] (including some for inscriptions that have now been destroyed) yet the use of these collections has been limited due to a variety of factors including the expense of travel and the difficulties of preservation. Consequently, Barmpoutis et al. sought to develop methods that could store and preserve squeezes and make them more accessible to a larger number of scholars.

The authors developed a framework that uses "3D reconstruction of inscriptions" and "statistical analysis of their reconstructed surfaces." They used a regular image scanner to scan squeezes from two different lighting directions and these images were then used in a "shape from shading technique in order to reconstruct in high resolution the original 3D surface." Barmpoutis et al. argued that the major contributions of their research were

[296] The focus on this section has been in looking at research that sought to provide better access to images of inscriptions (e.g. to enhance access to inscription text) rather than on virtual reconstruction of the monuments or other objects on which they are found, for some recent work in this area, please see (Remondino et al. 2009).
[297] http://bvreh.humanities.ox.ac.uk/VRE-SDM.html
[298] For an interesting digital collection of squeezes, see Ohio State University's Center for Palaeographical Studies, "Greek and Latin Inscriptions: Digital Squeezes" http://drc.ohiolink.edu/handle/2374.OX/106

threefold: 1) they had developed the first framework for converting and storing squeezes in 3D; 2) their research demonstrated how squeezes could be studied more effectively using different visualizations and such results could be more easily shared and distributed; 3) automated analysis of the squeezes produced results that would likely have been impossible to obtain with traditional techniques. Their framework was applied to five Ancient Greek inscribed fragments from Epidauros in southern Greece and they conducted experiments in surface recognition and statistical analysis. The ability to use different visualizations and 3D data they also proposed would support collaborative work and preservation:

> Rendering the inscriptions with different virtual illuminations and viewing angles makes the use of a squeeze more effective and allows the archaeologists to share digital copies of the squeezes without losing any information. Thus by using our proposed framework the digital libraries of scanned squeezes (regular 2D images) which are commonly used by archaeology scholars can easily be replaced by databases of 3D squeezes, without the need of any additional equipment (Barmpoutis et al. 2009).

In addition, the use of statistical analysis (such as automatically creating height-maps of the average letters) both replicated the results of individual scholars and also helped epigraphists to significantly speed up the process of analyzing individual letters, the variability of lettering schemes and evaluating their results.

Other advanced research in the imaging of inscriptions has investigated the automatic classification of Greek inscriptions according to the writer who carved them (Panagopoulos et al. 2008). One of the biggest challenges Panagopoulos et al. noted in studying inscriptions carved on stone is that they are unsigned, undated and have often been broken up and so are in various fragments. At the same time, they proposed that identifying a writer could be a crucial part of dating an inscription and thus setting it in its historical context. The major goals of their work were to objectively assign inscriptions to writers, to assist in writer identification where archaeological information and analysis had yielded no results, and to help resolve archaeological disputes regarding the dating of events. In sum, they reported that they hoped to "achieve writer identification employing only mathematical processing and pattern recognition methods applied to the letters carved in each inscription" (Panagopoulos et al. 2008).

One archaeologist worked with several computer scientists to evaluate the final methodology described in the paper. They obtained images of 24 inscriptions, segmented the images and extracted the contours of individual letters. Using mathematical processing they computed "platonic" prototypes for each alphabet symbol in each inscription. All inscriptions were then "compared pairwise by employing these ideal representations and the individual letter realizations." Panagopoulos et al. then used several statistical techniques to reject the "hypothesis that two inscriptions are carved by the same writer" and finally computed maximum likelihood considerations in order to definitively attribute inscriptions in their collections to their individual writers. To evaluate their framework, they used it to automatically attribute 24 inscriptions from Athens and successfully attributed these inscriptions to six different identified "hands" and matched the expert opinions of epigraphists. One particular strength of their process, the authors concluded, was that it required no training data, but they also hypothesized that a greater mass of inscription data on which to test their system would help them to greatly improve their accuracy rate.

Manuscript Studies

Any discussion of manuscripts quickly leads to the examination of many challenges found across classical disciplines such as the creation of digital editions, the complications of palaeographic studies, and how to design a digital collection of manuscripts that supports researchers considering codicological, [299] historical or philological questions. Manuscripts are one of the most complicated and highly used artifacts across disciplines. The data richness of manuscripts, according to Choudhury and Stinson in their analysis of the commonalities

[299] Codicology has been defined as "the study of the physical structure of books, which, when used in conjunction with palaeography, reveals a great deal about the date, place of origin, and subsequent history of a particular codex. The term was first used in conjunction with listing texts in catalogue form, but later in the 20th century came to be associated primarily with the structural aspects of manuscript production, which had been studied in a coherent fashion since the late 19th century." Timothy Hunter "Codicology." *The Oxford Companion to Western Art*. Ed. Hugh Brigstocke. Oxford University Press, 2001. Oxford Reference Online. Oxford University Press. Tufts University. 27 April 2010 <http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t118.e581>

between creating an infrastructure for a manuscript digital library and for a massive dataset in physics, makes them an intricate but important source for humanities data:

> Manuscripts, so evidently data-rich in the era in which they were created, today retain their former value and meaning while they inspire a new generation of humanists to create new sets of data. This includes the metadata needed to encode, organize, and understand the texts, annotations, and the visual art embodied in the manuscripts. Not only does this demonstrate the parallel need for data curation and preservation in the humanities and the sciences (for at the level of storage infrastructure, a byte is a byte and a terabyte a terabyte) but it underscores the fact that there is an increasing convergence of what it is that is analyzed by humanities scholars and scientists: data (Choudhury and Stinson 2007).

The authors noted that manuscripts represented the richest sets of data for their day because they integrated texts and images, included user annotations, as well as vast numbers of intertextual allusions and references.

While not specifically a "discipline" of classics, manuscript studies informs the work of many classical disciplines. The majority of classical texts—whether they are studied for philological analysis or as a source of ancient history—that form the basis for the modern critical editions upon which many scholars rely, are based off of medieval manuscripts. In addition, as was seen in the section on digital editions and textual criticism, access to images of manuscripts and transcriptions was cited as an essential component of cyberinfrastructure for classics. Thus this special subsection has been created to address some of the challenges of creating digital libraries of manuscripts and to examine individual research projects in detail. Some of the projects discussed here have received fuller treatment in other disciplinary sections of the paper.

Digital Libraries of Manuscripts

The last twenty years has seen a voluminous growth in the number of both digital images and transcriptions for manuscripts that have become available online. As indicated by the Catalogue of Digitized Manuscripts, there are both large collections of digital manuscripts at single institutions [300] and many individual manuscripts that have been digitized by individual libraries, museums or cultural organizations. [301]

One of the largest exemplary collections is "Medieval Illuminated Manuscripts" [302] a website that has been provided by the Koninklijke Bibliotheek and the Museum Meermanno-Westreenianum (Netherlands). This website serves as an extensive database of research information about illuminated medieval manuscripts. [303] Over 10,000 digital images of decorations taken from manuscripts are provided, and they may be browsed by subject matter (in English, German or French) using the ICONCLASS classification system that was created for the classification of art and iconography. For example, a user may choose "Classical Mythology and Ancient History" and then choose "Classical History" that then takes them to a final selection of options such as "female persons from classical history," selecting one of these options then takes the user to a list of manuscript images (with high resolution and zoomable images available) where picking an individual image also provides the user with a full manuscript description and a bibliography of the manuscript. A searchable database is also provided.

Two different projects, the Digital Scriptorium [304] and Manuscriptorium, [305] have been created to try and bring together large numbers of digital manuscripts online, or essentially to create virtual libraries of digital manuscripts. [306] Each has taken a different approach to this common problem.

The Digital Scriptorium provides access to a online image database of medieval and Renaissance manuscripts from almost 30 libraries and currently includes records for 5300 manuscripts and 24,300 images. This collection can be browsed by location, shelfmark, author, title, scribe, artist and language (including 58 Greek manuscripts). Each manuscript record includes an extensive bibliographic and physical description, links to

---

[300] For example, see "Digital Medieval Manuscripts" at Houghton Library, Harvard University, http://hcl.harvard.edu/libraries/houghton/collections/early_manuscripts/

[301] http://manuscripts.cmrs.ucla.edu/languages_list.php

[302] http://www.kb.nl/manuscripts/

[303] Some recent research has also explored innovative approaches to supporting more effective scholarly use of illuminated manuscripts through the development of user annotation tools and a linking taxonomy, see for example (Agosti et al. 2005).

[304] http://www.scriptorium.columbia.edu/

[305] http://beta.manuscriptorium.com/

[306] Other approaches have also explored developing large data grids or digital infrastructures for manuscripts, see (Calanducci et al. 2009).

individual manuscript pages images (thumbnail, small, medium, large), [307] and links to the fully digitized manuscript at its home institution (where available). Several types of searching are available including a basic search, a shelfmark search and an advanced search where a user can enter multiple keywords (to search the fields: shelfmark, author, title, docket, language, provenance, binding, caption) with limits by date, decoration, country of origin, and current location.

A brief overview of the Digital Scriptorium (DS) and its future has been provided by Consuelo Dutschke (Dutschke 2008). She articulated how the creation of the DS had made the work of text editors in assembling a body of evidence much simpler, and that libraries that had chosen to participate had also made the job of future editors far easier for it provides a single point of access to the indexed holdings of multiple libraries. She also observed that many libraries that had chosen to participate in DS had consequently made a much greater effort to identify their own collections. Even more importantly, however, the DS can help editors gain a more complete understanding of the context of the manuscripts with which they work:

> DS also serves the cause of the editor in allowing him a first glimpse of the world that a given manuscript occupies: the other texts with which it circulates; the miniatures, if any, which always imply interpretation; the level of expense that went into its production; early and late owners with their notes and their bindings, each bringing a historical glimpse of that manuscript's value – both semantic and financial – to the whole. Leonard Boyle reminds us that no text exists without its physical means of transmission…. and DS significantly aids the editor in building an understanding of the physical and intellectual environment of the chosen text (Dutschke 2008).

Dutschke asserted that an editors' understanding would also grow as they could examine other manuscripts of the same text or even other manuscripts of different texts but of a similar place and date of origin. The DS provides access to only some images of manuscripts (an average of six images per codex) as the costs of full digitization were prohibitive in many cases. Nonetheless it serves as an important discovery tool for widely scattered collections, Dutschke maintained, since for most researchers it simply matters if a library has the particular text, author, scribe or artist that they are researching.

The DS began in 1997 and first established standards for bibliographic data collection and photographic capture of manuscripts, standards that are iteratively updated. The existence of such standards along with documentation has also made it easier for potential collaborators to determine whether they wish to join the DS. These documentation and standards have also proved a crucial component of technical sustainability according to Dutschke. Nonetheless, the other critical element of sustainability, she pointed out is financial, and the DS is currently taking steps to ensure the survival of their digital program for the indefinite future. Part of any financial sustainability plan, Dutschke explained, was a concrete specification of what is required to keep an organization running as well as keep down future costs. Some key elements she listed included documentation, technological transparency, simplicity, and also sensible file naming. As Dutschke explained, "there is an unfortunate tendency to want to make the file name carry verbal meaning, to allow humans to understand how it refers to the real-life object" (Dutschke 2008). Yet she argued that this was unnecessary as the cataloging for files occurs elsewhere in tables of the database so such information need not be encoded in the file name. Simple and transparent file names, she insisted helped limit future costs of having to update invalid semantic values or fixing typing errors.

While the database used for data entry and collection is currently Microsoft Access, Dutschke also pointed out that on a regular basis every DS partner exports its own collection specific information into XML and then forwards that XML data to the central DS organization. "It is on the XML-encoded data that technology experts write the applications that make the data useful to scholars," Dutschke reported, "via meshing the data from multiple partners, searching it, retrieving it, displaying it." In addition, because XML is both non-proprietary and platform independent they have also chosen to use it for "data transport, long-term storage and manipulation." Cayless et al. (2009) have also argued for the use of XML as a long-term preservation format for digitally encoded epigraphic data. In addition, two other long-term preservation challenges that were also identified by Dutschke were the challenges of mass storage and the security of the files.

---

[307] For example, a large image of a manuscript page of Hero of Alexandria's *Geometrica*, http://www.columbia.edu/cgi-bin/dlo?obj=ds.Columbia-NY.NNC-RBML.6869&size=large

While technical interoperability and financial sustainability were two key components of the long-term preservation of any digital project, Dutschke ultimately concluded that the most important questions were political, or in other words, were the DS partners committed to its long-term survival and did the larger user community value it. In order to stabilize the DS consortium, the DS has created a governing body that has developed policies for the daily management of decision-making. They also conducted a user survey to which 200 people responded and 43 of which agreed to be interviewed in detail, with the major conclusion of this survey being the unrelenting demand for more content. This led Dutschke to offer the important insight that digital projects need to understand both their *current* and *future* user demands, ultimately positing that "the will to sustainability lies not only within the project and its creators/partners; it also lies with its users."

One last critical issue raised by Dutschke addressed not only the needs of the DS but of the digital humanities as a whole to develop a greater understanding of the costs and needs of cyberinfrastructure:

> It's not simply that digital projects cost money; all human endeavour falls into that category. It's that digital projects remain so new to us that we, as a nation and even as a world-wide community of scholars working in the humanities, haven't fully understood the costs nor factored them out across appropriate bodies. The steps DS has taken towards a more reliable ad efficient technology, and the steps it has not taken reflect growth and uncertainty in the field overall. DS and the digital world as a community still lack a cyberinfrastructure not simply in terms of hardware or software, but even more importantly as a shared and recognized expertise and mode of operation (Dutschke 2008).

Throughout this review, the uncertainty regarding the way forward for a digital infrastructure on both individual and cross-disciplinary levels was frequently discussed. While technological issues and financial concerns were often raised, Dutschke also broached the important question of the lack of shared expertise and business models in terms of understanding how best to move forwards towards building a humanities cyberinfrastructure.

The other major virtual library of manuscripts is the Manuscriptorium, a project that according to its website is seeking to create "a virtual research environment providing access to all existing digital documents in the sphere of historic book resources (manuscripts, incunabula, early printed books, maps, charters and other types of documents)." [308] Manuscriptorium provides access to more than 5 million digital images from dozens of European as well as several Asian libraries and museums. Extensive multilingual access is provided to this collection (including English, French, German and Spanish). In addition to multilingual searches, a translation tool provided by Systran can also be used to translate manuscript descriptions from one language to another. This rich multilingual environment includes both modern and ancient languages and this has had led to complicated transcription issues since the collection includes Old Slavonic, Greek, Arabic, Persian and various Indian languages.

The Manuscriptorium collection can be searched by document identification or document origin, and also provides both an easy and advanced search interface. The "easy search" allows a user to search for documents by location, keyword, timeframe, responsible person or associated name, or for those documents with a digital facsimile, full text transcription, edition or transliteration. The "advanced search" offers multiple keyword entry in the fields (shelf-mark, text anywhere, country, settlement, and library). A user search brings up a list of relevant manuscripts and each manuscript description includes full bibliographic information, physical description and a link to a full digital facsimile (when available). Opening up a digital facsimile then launches a separate image viewer for examining the images of the facsimile that are available.

A recent article by Knoll et al. (2009) has provided some further explanation of the history, technical design and goals of the Manuscriptorium project. It first began as the Czech Manuscriptorium Digital Library in 2002 and through the ENRICH project [309] was expanded to provide seamless access to data about and digital images of manuscripts from numerous European institutions. Manuscriptorium supports harvesting via OAI for existing digital collections of manuscripts and has also created tools to allow participating organizations that simply wish to create their digital library as part of the larger Manuscriptorium to create compliant data. The development of

---

[308] http://beta.manuscriptorium.com/

[309] The recently concluded ENRICH project (funded under the EU eContent + progamme) provided funding to expand Manuscriptorium to serve as a digital library platform "to create seamless access to distributed information about manuscripts and rare old printed books." http://enrich.manuscriptorium.com/

Manuscriptorium has involved the creation of both technical and legal agreements that have also evolved over time. The early standard used in manuscript description was the MASTER DTD [310] and participating in the project also required that partners provide detailed technical descriptions about their manuscript images as well as a "framework for mapping the document structure with references to images."

Under the auspices of the ENRICH program, the Manuscriptorium project realized that a new more robust DTD would be necessary. This was a rather complicated process as they sought to harvest data created from two very different approaches to manuscript description, that of the library community (MARC) and that of researchers and text encoders (TEI):

> In the so-called catalogue or bibliographic segment, there are certain description granularity problems when converting metadata from MASTER (TEI P.4) to MARC, while vice versa is not problematic. On the other hand, TEI offers much more flexibility and analytical depth even in the description segment and, furthermore, being a part of a complex document format, it also provides space for structural mapping (Knoll et al 2009).

Since two major goals of the ENRICH project were to support both data interchange and sharing as well as data storage, Oxford University Computer Services led the development of both a new TEI P5 compliant DTD and schema. [311] It was decided that TEI supported not only all the requirements of manuscript description but also provided a common format for structural mapping and would be able to accommodate all incoming levels of manuscript granularity. All existing documents in the digital library thus had to be migrated from the original masterx.dtd to the new one created, and all documents that are harvested or added directly must conform to it as well.

The basic image access format used within Manuscriptorium is JPEG but they also support GIF and PNG. Nonetheless, one major challenge in data integration Knoll et al. reported was that many individual libraries chose to provide access to images of their manuscripts through multi-page image files such as PDF or DjVu rather than through XML based structural mapping. Since this kind of access did not support the manipulation of manuscript pages as individual digital objects, all participating libraries were required to convert such files into individual JPEG images for each manuscript page. As Knoll et al. explained:

> The goal of Manuscriptorium is to use its own interface for representation of any document from any partner digital library or repository. Thus, the central database must contain not only descriptions of such documents – manuscripts or rare old prints – but also their structural maps with references (individual URLs) to concrete image files. As the user wishes to consult the concrete files, they are called into the uniform Manuscriptorium viewer from anywhere they are so that the user enjoys seamless access without navigating to remote digital libraries or presentations (Knoll et al. 2009).

The ability to provide a seamless point of access to distributed collections of digital objects or in effect to create a virtual user experience was also cited as important by the CLAROS, LaQuAT, and TextGrid projects. In the end, an ideal partner for Manuscriptorium as described by Knoll et al. is one whose collection can be harvested by OAI and where each manuscript profile contains both a descriptive record and a structural map with links to any images. While harvesting and transforming descriptive records was fairly simple, Knoll et al. also reported that harvesting structural mappings was far more problematic. Two specific tools have also been created to allow content providers to easily create the structured digital documents required by Manuscriptorium. The first tool M-Tool supports both the manual entry and creation of new metadata so it can be used to either create new manuscript records within Manuscriptorium or to edit existing ones for import. The second tool M-Can has been specifically created for uploading and evaluating manuscript records (Marek 2009).

One of the most innovative features of Manuscriptorium is that it supports the creation of personal digital libraries. Users who register (as well as content providers) can "build their own virtual libraries from the aggregated content" and thus organize content into static personal collections or dynamic collections (based off of a query so the collection automatically updates based on your query terms). Even more importantly, users can create "virtual documents" that can be shared with other users. These documents can be created through the use of the M-Tool application. These virtual documents are particularly interesting for they can be composed of

[310] http://digit.nkp.cz/MMSB/1.1/msnkaip.xsd
[311] http://tei.oucs.ox.ac.uk/ENRICH/ODD/RomaResults/enrich.dtd and http://tei.oucs.ox.ac.uk/ENRICH/ODD/RomaResults/enrich.xsd

parts of different physical documents (individual page images from *different* manuscripts can be saved with notes from the user), and in one example they give, "interesting illuminations from manuscripts of a certain period" could be selected and "bound" into a new virtual document." [312] Images from other external URLs can also be inserted into these virtual documents. This level of personalization is extremely useful and rarely found among most digital projects. Moreover, the ability to create virtual manuscripts that contain page images of interest from various manuscripts will likely support interesting new research.

As indicated by the projects briefly surveyed here, there are a wealth of digital manuscript resources available online, and the next section will look at some of the specific challenges of working with complicated individual manuscripts.

Digital Challenges of Individual Manuscripts and Manuscript Collections

This review has already briefly explored some of the challenges of manuscript digitization in terms of advanced document recognition and research projects such as the work of EDUCE with the Homer Multitext as well as research conducted on the Archimedes Palimpsest.

This section will briefly examine another perspective, the challenges of creating metadata (e.g. linking transcriptions, translations and images) to manage highly complicated individual digital manuscripts such as the Codex Sinaiticus [313] and the Archimedes Palimpsest and for digital collections of the multiple manuscripts of a single work such as in the Roman de La Rose Digital Library.

The Codex Sinaiticus is one major project that illustrates some of the challenges of creating a digital library of an individual manuscript, albeit one that exists in fragments in various collections. This manuscript was hand written over 1600 years ago and contains a copy of the Christian Bible in Greek, including the oldest complete copy of the New Testament. This text that has been heavily corrected over the centuries and is of critical importance not just for Biblical studies but also as the oldest "substantial book to survive antiquity" is a important source of study for the "history of the book." The original codex was distributed in unequal portions between London, Leipzig, Sinai and St. Petersburg and an international collaboration reunited the manuscript in digital form and has made it available online. A recent article by Dogan and Scharsky (2008) has provided some description of the technical and metadata processes involved in creating the digital edition of this codex that is available online. They stated that creation of the website involved physical description of the manuscript, translation of selected parts into different languages such as German and English, the creation of a Greek transcription and the digital imaging of the entire codex.

On the website, the user can choose to view the manuscript transcription by either "semantic layout" (view by Biblical verse) or by manuscript layout (view by page). An image of the codex is presented with both the Greek transcription and often a parallel translation of the verse in various languages (when available). The entire codex can also be searched (including the transcription or the translation) and a Greek "keyboard" is available to search in Greek. Dogan and Scharsky have also reported that multispectral images of the manuscript were also taken in order to "enable erased or hidden text to be discovered as well as codicological and palaeographical characteristics of the Codex to be fully analysed." Another major challenge they noted was that almost every page includes "corrections, re-corrections and insertions, many of considerable textual significance." One final goal of the project is to make available a fully searchable electronic transcription of both the main text and corrections. The project developed a specific schema based on the TEI to create a transcription that reflected both the Biblical structure (book, chapter, verse) and the physical structure (quire, folio, page, column) of the manuscript. Development of the website has also involved creating a specialized linkage system between image, transcription and translation.

While the advanced document recognition technology used with the Archimedes Palimpsest has been discussed previously, the metadata and linking strategies used to link manuscript metadata, images and transcriptions that

[312] For more on the creation of personal collections and virtual documents, see http://beta.manuscriptorium.com/apps/main/docs/mns_20_pdlib_help_eng.pdf
[313] http://www.codexsinaiticus.org/en/

were developed also merit some further discussion. Two recent articles by Doug Emery and Michael B. Toth (Emery and Toth 2009, Toth and Emery 2008) have described this process in detail. The creation of the Archimedes Palimpsest Digital product, which released one terabyte of integrated image and transcription data, required the spatial linking of registered images for each leaf "to diplomatic transcriptions that scholars initially created in various nonstandard formats, with associated standardized metadata" (Emery and Toth 2009). The transcription encoding built off of previous work conducted by the Homer Multitext project, and Emery and Toth noted that standardized metadata was critical for three purposes: "1) access to and integration of images for digital processing and enhancement, 2) management of transcriptions from those images, and 3) linkage of the images with the transcriptions."

The authors also described how the great disciplinary variety of scholars working on the palimpsest from students of Ancient Greek to those exploring the history of science necessitated the ability to capture data from a range of scholars in a standard digital format. This necessity led to a "Transcription Integration Plan" that incorporated Unicode, Dublin Core and the TEI. They explained that they chose Dublin Core as their major integration standard for digital images and transcriptions because it would allow for "hosting and integration of this data set and other cultural works across service providers, libraries and cultural institutions"(Toth and Emery 2008). While they utilized the "Identification," "Data Type," and "Data Content" elements from the Dublin Core element set, they also needed to extend this standard with elements such as "Spatial Data Reference" drawn from the Federal Geographic Data Committee *Content Standard for Digital Geospatial Metadata*.

Emery and Toth (2009) argued that one of the guiding principles both behind their choice of common standards and emphasis on the importance of integrating data and metadata was the need to create a digital archive for both today and the distant future. The data set they created thus also follows the principles of the Open Archival Information System (OAIS) [314] In their data set, every image bears all relevant metadata in its header and each image file or folio directory serves as a self-contained preservation unit that includes all the images of a given folio side, XMP metadata files, checksum data and the spatially mapped TEI-XML transcriptions. In addition, the project developed its own Archimedes Palimpsest Metadata Standard that "provides a metadata structure specifically geared to relating all images of a folio side in a single multi- or hyper-spectral data "cube"" (Emery and Toth 2009). Because each image has its own embedded metadata the images can either stand alone or be related to other members of the same cube. Finally, over 140 of the 180 folio sides include a transcription and the lines in these transcriptions are mapped to rectangular regions in the folio images using the TEI <facsimile> element. This mapping serves two useful purposes, first, it allows the digital transcriptions to provide "machine readable content," and second, it allows easy movement between the transcription and the image.

In addition to the challenges presented by individual manuscripts, other digital projects have explored the challenges of managing multiple manuscripts of the same text. The Roman de La Rose [315] Digital Library (RRDL), a joint project of the Sheridan Libraries of Johns Hopkins University and the Bibliothèque Nationale de France (BnF), seeks to ultimately provide access to digital surrogates of all of the manuscripts (over 300) containing the *Roman de la Rose* poem. The creation of this digital library was supported by the Mellon Foundation and by the end of 2009 the website provided access to digital surrogates of roughly 130 manuscripts through either a French or English interface. The website includes a list of all extant manuscripts as well as a collection spreadsheet that can be sorted by various columns if a user wants to sort manuscripts alphabetically or by number of illustrations. Clicking on any individual manuscript name links to a full codicological description [316] that also includes a link to the digitized manuscript. Individual manuscripts can be read page-by page in a special viewer with a variety of other viewing options such as (full screen) or zooming in on the individual pages. [317] For several manuscripts, a transcription can also be viewed on screen at the same time as

[314] For more on this ISO standard, see http://public.ccsds.org/publications/archive/650x0b1.pdf
[315] http://romandelarose.org/#home
[316] Manuscript descriptions have been encoded in TEI P5 (Stinson 2009).
[317] http://romandelarose.org/#read;Douce195.156v.tif

some individual manuscript pages. Individual and citable URL's are provided for the codicological descriptions [318] and for two-page views of the manuscripts within the special viewing application. [319]

In addition to choosing manuscripts from the collection spreadsheet, individual manuscripts can also be chosen by browsing the collection by repository, common name, current location, date, origin, type, number of illustrations or folios, and availability of transcription. Once a manuscript has been selected, a user can choose to examine the codicological description, to view it in the "page turner" application described above or to simply browse the images in a special viewer. Each manuscript also includes a full bibliography. Both a basic keyword and advanced search feature are available, and the advanced search allows for multiple keyword searching within various fields (lines of verse, rubric, illustration title, narrative sections, etc.)

With such a large number of digital surrogates available, one of the most significant opportunities presented by the RRDL is the possibility of "cross manuscript comparative study." In order to facilitate this, the creators of this collection found it necessary to create a new text organizational structure called narrative sections as explained on the website: [320]

> Citation practice for the Roman de la Rose and most medieval texts has traditionally referenced the currently accepted critical editions. Yet this scholarly protocol inhibits the cross-manuscript comparative study that the Roman de la Rose Digital Library promotes. Since the number of lines for the work varies from one manuscript to another, depending on interpolations or excisions, the narrative mapping of the Roman de la Rose divides the text into reading segments instead of lines. This means that comparable passages across different manuscript can be readily locatable, while number of lines for each section facilitate tracking variations in section length from one exemplar to another. The narrative mapping protocol borrows from that used for classical texts, where one cites not a page number or a given edition or translation but a segment of the text.

The narrative mapping was largely generated algorithmically but should apparently be accurate to within one or two columns of texts. By selecting a narrative section, the user can then be taken to a list of image sections for the different manuscripts that contain that section so they can compare the individual manuscript sections themselves. The need to create a new canonical text structure independent of particular scholarly editions or conventions in order to facilitate the citation, navigation and use of manuscripts in a digital environment was an issue also articulated by the creators of the CTS in terms of classical texts, indicating that there are many similar digital challenges to be resolved across disciplines.

Another significant manuscript project is that of Parker on the Web, [321] a multi-year project of Corpus Christi College, Stanford University Libraries and Cambridge University Library, to create high-resolution digital images of almost all the manuscripts in the Parker Library. This project has built an "interactive web application" to allow users to examine manuscripts within the "context of supporting descriptive material and bibliography." There are over 550 manuscripts described on this site and almost all of them were numbered and catalogued by M.R. James in his 1912 publication. The online collection also includes some volumes it received after the publication of the James catalogue. Limited free access to the collection is provided but full access is only available by subscription.

The digitization of these two projects and their consequent effects upon manuscript studies and codicology in particular have been explored by (Stinson 2009). Stinson noted that in the RRDL all digital surrogates were connected to codicological descriptions since many important features of manuscripts as physical books can be lost when represented in digital form. In addition, no comprehensive catalog or reference work existed that contained either descriptions or a full list of all the Rose manuscripts, so the project team wrote many of these descriptions themselves. In contrast, Parker on the Web was able to create marked up descriptions of the entries for manuscripts in the M.R. James catalogue. This very process however, led them to some important conclusions:

> Yet in marking up both sets of descriptions—one custom made for the web, the other a digitized version of a printed reference work—for inclusion in digital libraries, and in designing and implementing interfaces for accessing XML-encoded descriptions and the surrogates to

[318] http://romandelarose.org/#book;SeldenSupra57
[319] http://romandelarose.org/#read;SeldenSupra57.013r.tif
[320] http://romandelarose.org/#sections
[321] http://parkerweb.stanford.edu/parker/actions/page.do?forward=home

which they are linked, it has become apparent that in digital form the relationship of codicological descriptions to the books they describe has, like the relationships of critical editions to the texts they document and represent, undergone fundamental change (Stinson 2009).

The digital environment has changed codicological descriptions in three major ways, according to Stinson: 1) new purposes and uses have been discovered for these descriptions, particularly in terms of their specific and technical language 2) the relationship between a codicological description and codex has moved from one-to-one to a one-to-many relationship between "codices, descriptions, metadata and digital images" and 3) where once books were used to study other books, digital tools are now being used to represent and analyze books. These insights also emphasize the larger realization that when printed reference works are digitized—particularly when the knowledge they contain is marked up in a meaningful way—they can take on whole new roles in the digital world. [322]

Stinson described how codicological descriptions were typically created by experts using a formalized vocabulary to summarize dates, origins, owners, contents of books, among other items and that these descriptions were used either by visitors to a library who wanted to use a manuscript or for scholars studying the manuscript remotely. In digital libraries, however, Stinson argued that digital images of codices serve as the "machine readable forms of the original artifacts" and that "XML encoded codicological descriptions are the secondary information used to describe, analyze and interpret these artifacts." While codicological descriptions are still needed for the dissemination of specialized knowledge (such as for the palaeographical and literary histories of individual manuscripts), Stinson argued that their purpose of physical description and providing information to remote scholars has evolved in a digital environment. Physical description can still be important since digital repositories often misname files rather than making mistakes in foliation and pagination, and "a break in a digital codex might as easily be the result of a lost file as a lost leaf in the physical book it represents." Even more importantly, the extensive descriptive information once intended to aid remote scholars now provides new means for "sorting, classifying and comparing collections of manuscripts." Although 17,000 word transcriptions Stinson admits can't easily be put into a relational database, specific information can be extracted from them:

…the precision and specificity of the language of codicological descriptions, developed to convey a substantial amount of information in a small space (a necessity in print reference works if one wishes to avoid prohibitive cost and unwieldy volumes) now facilitates databases that provide highly flexible, searchable, and sortable relationships between the original artifacts (Stinson 2009).

In fact, as described above, the RRDL provides access to a complete database created from much of the codicological information and it can be viewed online, downloaded as a spreadsheet, or used to search or sort this information "across the entire corpus of manuscript descriptions." [323]

The second major change Stinson listed was how the new many-to-many relationship between a codicological description its codex and the images that constitute the digital surrogate has created a new series of complex relationships that must be represented in a digital library environment. Codicological descriptions in a digital environment can be hyperlinked not just to digital images of the codex itself but to digitized items listed in its bibliography, biographies of illustrators, and indeed to any related scholarly information that is available online. These codicological descriptions then not only continue to serve as guides to the printed codices and their digital surrogates, but because they have been marked up in XML with defined data categories, can be used to create databases and in combination serve "as a large searchable "meta-manuscript" that contains combined data from numerous physical codices and thousands of digital images" (Stinson 2009).

The final insight offered by Stinson underscored how the print environment has expanded the potential for printed reference works that were once used solely to study other books, for now these reference works can be turned into digital tools that can then provide much more sophisticated opportunities for analysis. He asserted that printed codicological descriptions, such as those found in the James catalogue, suffer from the same challenges of many printed critical editions, in particular, they include a large number of "abbreviated and coded

[322] For further consideration of how digitized historical reference works can be utilized in new ways see (Crane and Jones 2006) and (Gelernter and Lesk 2008).
[323] http://romandelarose.org/#data

forms" known only to experts. [324] Such abbreviated forms were used as space saving devices that are no longer necessary in a digital environment and Stinson thus insists that much of the data embedded in codicological descriptions "lies latent" until it is "unleashed" by digitization. At the same time, he argued that the digitization of both manuscripts and their codicological descriptions offered a new opportunity to move beyond simple digital incunabula: [325]

> The rubrication, historiated initials, and foliated borders of incunables remind us that in the early days of print the concept of what a book should be was dominated by the manuscript codex. During recent centuries, the opposite is true; descriptions of manuscript books bear witness to the dominance of printing in forming our collective notion of what a book should be…. As we seek to liberate our codicological descriptions from the constraints of "being compelled to operate in a bookish format," we should also bear in mind the opportunity to correct the assumption that such books operate—and should be described—in parallel with printed books. Both our tools and our mindsets need to be liberated from print if we are to achieve accurate representations of artifacts that were produced before the advent of printing (Stinson 2009).

The need to go beyond traditional printed models and use the digital environment to both more accurately represent cultural objects and artifacts and to "unleash" their latent semantic potential, whether they are primary texts, archaeological monuments in ruins, inscriptions on stone or medieval manuscripts is a theme that has been throughout this review.

Digital Manuscripts, Infrastructure and Automatic Linking Technologies

As illustrated by the previous sections, any digital infrastructure designed for manuscripts will need to address the complicated nature of manuscripts as both physical and digital objects, support a range of scholarly uses, and provide effective access to all of the data created in their digitization (e.g. digital images, diplomatic transcriptions, TEI-XML editions, scholarly annotations). One challenge is that while there are many images of digital manuscripts available online, many are only viewable through special image-viewers and thus often do not have stable URLs that can be cited. Furthermore, many of these digitized manuscripts do not have individual URL's for each page so linking to a specific page let alone an individual line or word is impossible. Similarly, it is often difficult to determine if a digitized manuscript has a transcription available and even if one has been created it is often even more difficult for a user to gain access to it.

Two projects that are seeking to create some of the necessary infrastructure for manuscripts that will address some of these issues are the Interedition project and the Virtual Manuscript Room (VMR). The Interedition project is focused in particular on developing a "supranational networked infrastructure for digital scholarly editing and analysis." As the creation of digital editions particularly in classical and medieval studies typically involves the use of multiple manuscripts, their draft architecture addresses the need to represent multiple manuscripts and link to both individual manuscript images and transcriptions. [326]

While Interedition is focused on the larger infrastructure required for digital editions, the VMR, [327] which is currently in its first phase, is concentrating on providing advanced access to an important collection of manuscripts. Currently they have provided access to fully digitized manuscripts from the Mingana Collection of Middle Eastern Manuscripts at the University of Birmingham. Each digitized manuscript includes high-resolution images of each page and descriptions from both the printed catalogue and the special collections department that holds them. In their next phase, they will add even more content, including 50,000 digital manuscript images, 500 manuscript descriptions and 1000 transcription pages. Even more importantly, however, the next phase of the VMR's work will also involve the development of a framework for digital manuscripts that will:

> …bring together digital resources related to manuscript materials (digital images, descriptions and other metadata, transcripts) in an environment which will permit libraries to add images, scholars to add and edit metadata and transcripts online, and users to access material…. (http://vmr.bham.ac.uk/about/)

---

[324] Both Bodard (2008) and Roueché (2009) have observed similar opportunities of expanding specialist abbreviations in the digital editions of inscriptions, and Rydberg-Cox (2009) has also described the challenges of digitizing abbreviated texts found within incunabula.
[325] The need to move beyond "digital incunabula" has been articulated in (Crane et al. 2006).
[326] http://www.interedition.eu/wiki/index.php/WG2:Architecture
[327] http://vmr.bham.ac.uk/

As part of this phase, the VMR at the University of Birmingham [328] also plans to join together with a parallel VMR being built at the University of Münster Germany and provide seamless access to both collections. Four key features distinguish this work from previous manuscript digitization projects: 1) it is designed around granular metadata, so instead of simply presenting metadata records for whole manuscripts, records are provided for each page image, for the transcription of the text on that page, and for specifying what text is on that page, 2) "the metadata states the exact resource type associated with the URL specified in each record" (e.g. if a text file is in XML and what schema has been used), 3) all VMR materials will be stored in Birmingham's institutional repository and be accessible through the library online public access catalog (OPAC), and 4) the VMR will support full *reuse* of its materials not just access to them.

This fourth feature is perhaps most unique, for as seen by the survey of projects in this section, the focus of much manuscript digitization work has often been on supporting the *discovery* of digital manuscripts for use online rather than on the ability for scholars to get access to the raw digital materials and use them in their own projects. The VMR plans to provide access to all the metadata they create through a syndicated RSS feed so that users can create their own interfaces to VMR data. In addition, they also plan to allow other users to add material to the VMR by creating a "metadata record for the resource following VMR protocols" and then add it to the RSS feed of any VMR project. The importance of supporting new collaboration models that allow many individuals to potentially contribute related digital manuscript resources has also been discussed in (Robinson 2009, Robinson 2010).

While there is an increasing amount of metadata about manuscripts as well as digital images and transcriptions of manuscripts that have become available online, there are still few easy ways to link between them if they exist in different collections. A related problem is the limited ability to at least partially *automate* the linking of manuscript images with their transcriptions, even if both are known to exist. Arianna Ciula has argued that the work of palaeographers would greatly benefit from descriptive encoding or technology that supported more sophisticated linking between images and texts, particularly "the possibility to export the association between descriptions of specific palaeographical properties and the coordinates within a manuscript image in a standard format such as the encoding proposed by the TEI facsimile module or SVG" (Ciula 2009).

Recently Hugh Cayless has developed a series of tools and techniques to assist in this process that have been grouped under the name img2XML [329] and have been described in detail in Cayless (2008, 2009). As has been previously discussed by (Monella 2008, Boschetti 2009), Cayless noted that manuscript transcriptions are typically published in one of two formats, either as a critical edition where the editors' comments are included as an integral part of the text or as a diplomatic transcription that tries to "faithfully reproduce the text" (Cayless 2009). While TEI allows the production of both types of transcriptions from the same marked up text, Cayless argued that the next important step is to automatically link such transcriptions to their page images. While many systems link manuscript images and transcriptions on the page level, [330] the work of Cayless sought to support even more granular linking, such as at the level of individual lines or even words.

Cayless thus developed a method for generating a "Scalable Vector Graphics (SVG) [331] representation of the text in an image of a manuscript" (Cayless 2009). This work was inspired by experiments conducted using the OpenLayers [332] Javascript library by Tom Elliott and Sean Gillies to trace the text on a sample inscription [333] and Cayless sought to create a "toolchain" that used only open source software. To begin with Cayless converted JPEG images of manuscripts into a bitmap format using ImageMagick [334] and then used an open source tool

[328] The VMR at Birmingham has been funded by JISC and is being created by the Institute of Textual Scholarship and Electronic Editing (ITSEE), http://www.itsee.bham.ac.uk/
[329] http://github.com/hcayless/img2xml
[330] One such system is EPPT (discussed earlier in this paper), and another tool listed by Cayless is the Image Markup Tool (IMT), http://www.tapor.uvic.ca/%7Emholmes/image_markup/index.php, which allows a user to annotate "rectangular sections of an image" by using a drawing tool with which they can first "draw shape overlays on an image" and then these overlays can then be linked to "text annotations entered by the user." (Cayless 2009).
[331] SVG "is a language for describing two-dimensional graphics and graphical applications in XML." http://www.w3.org/Graphics/SVG/
[332] http://trac.openlayers.org/wiki/Release/2.6/Notes
[333] http://sgillies.net/blog/691/digitizing-ancient-inscriptions-with-openlayers
[334] http://www.imagemagick.org/

called Potrace [335] to convert the bitmap to SVG. The SVG conversion process required some manual intervention and an SVG editor called InkScape was used to cleanup the resulting SVG files. The resulting SVG documents were then analyzed using a Python script that attempted to "detect lines in the image and organize paths within those lines into groups within the document" (Cayless 2008).

After the text image within a larger manuscript page image was converted into SVG paths, these paths could be grouped within the document to mark the words therein and these groups could then be linked using various methods to tokenized versions of the transcriptions (Cayless 2009). Cayless then used the OpenLayers library to simultaneously display the linked manuscript image and TEI transcription, for importantly, OpenLayers "allows the insertion of a single image as a base layer (though it supports tiled images as well), so it is quite simple to insert a page image into it" (Cayless 2008). This initial system also required the addition of several functions to the OpenLayers library, particularly the ability to support "paths and groups of paths." Ultimately, Cayless reported that:

> The experiments outlined above prove that it is feasible to go from a page image with a TEI-based transcription to an online display in which the image can be panned and zoomed, and the text on the page can be linked to the transcription (and vice-versa). The steps in the process that have not yet been fully automated are the selection of a black/white cutoff for the page image, the decision of what percentage of vertical overlap to use in recognizing that two paths are members of the same line, and the need for line beginning (<lb/>) tags to be inserted into the TEI transcription (Cayless 2008).

While automatic analysis of the SVG output has supported the detection of lines of text in page images, work continues in order to allow the automatic detection of *words* or other features in the image. Cayless also concluded that this research raised two issues. To begin with, he stated that further research would need to consider what structures (beyond lines) could be detected in a SVG document and how they could be linked to transcriptions. Secondly, TEI transcriptions often define document structure in a "semantic" rather than physical way, and even though line, word, and letter segments can be marked in TEI they often are not, and this fact makes it difficult if not impossible to automate the linking process. Cayless proposed that a standard would need to be developed for this type of linking.

Other experiments in automatic linking of images and transcriptions have been conducted by the TILE project. [336] This project seeks to build a new "web-based image markup tool" and is based on the existing code of the Ajax XML (AXE) image encoder. [337] In addition, it will be interoperable with both the EPPT and the IMT and be "capable of producing TEI-complaint XML for linking image to text." Similar to Cayless, they want to support linking beyond the page level, such as the ability, for example, "to link from a word in the edited text to its location in the image" or to "click an interesting area in the image to read an annotation" (Porter et al. 2009). While Porter et al. acknowledged that there were a number of tools [338] that allowed users to edit or display images within the larger context of creating digital editions, none of these tools contained all the functionality they desired.

Of all of the tools they mention, Porter et al. stated that only the IMT outputs complete and valid TEI P5 XML, but it also only runs on Windows machines. While TILE will interoperate with the "constrained IMT TEI format," it will also provide output in a variety of formats. A recent blog entry by Dorothy Porter listed these formats as including "any flavour" of TEI, METS [339] files, and output that is not in XML. "One result of this flexibility is that, again unlike the IMT, TILE will not be "plug and play", and processing of the output will be the responsibility of projects using the software," Porter acknowledged, "This will require a bit of work on the part of users. On the other hand, as a modular set of tools, TILE will be able to be incorporated into other digital

---

[335] http://potrace.sourceforge.net/

[336] This project's approach to digital editions was discussed earlier in this paper.

[337] http://mith.info/AXE/

[338] Among this list were Juxta (http://www.nines.org/tools/juxta.html), developed by the NINES project, which is typically used to compare two documents but also only connects images and text at the page level, and the Versioning Machine (http://v-machine.org/), a tool with some of the same basic functionality as Juxta, but again one that only supports the linking of texts and images at the page level.

[339] METS stands for "Metadata Encoding & Transmission Standard" and has been created by the Library of Congress "for encoding descriptive, administrative, and structural metadata regarding objects within a digital library"(http://www.loc.gov/standards/mets/). METS is expressed using XML and has been used by many digital library projects.

editing software suites that would otherwise have to design their own text-image linking functionality or go without" (Porter 2010).

Since AXE enabled the collaborative tagging of TEI texts, the association of XML with "time stamps in video or audio files" and the marking of image regions that could then be linked to external metadata, TILE will extend these functionalities. One significant issue with AXE was that while it did allow users to both annotate image regions and store those coordinates in a database, it did not provide any data analysis tools for this information. The most significant way in which TILE will extend AXE then is that it will support:

> Semi-automated creation of links between transcriptions and images of the materials from which the transcriptions were made. Using a form of optical character recognition, our software will recognize words in a page image and link them to a preexisting textual transcription (Porter et al. 2009).

As with the research of Cayless, the principal goal of this work is to be able to link manuscript transcriptions and images at the individual *word* level. Some other intended functionalities include image annotation with controlled vocabularies, the creation of editorial annotations, [340] and the creation of links between "different, non-contiguous areas of primary source images" such as captions and illustrations or "analogous texts across different manuscripts."

Numismatics

Numismatics has been defined as "the collection and study of money (and coins in particular)" [341] and is one of the most popular classics topics in terms of academic, commercial and enthusiast sites online. [342] In fact, according to Sebastian Heath (Heath 2010) any discussion of numismatics online must consider commercial and enthusiast websites as they often provide the most information online regarding coins, and he also asserted that some commercial enterprises were more open with their data than academic ones. In addition, a brief if informal survey conducted by Heath in terms of the findability of academic numismatics sites using Google illustrated that "commercial and personal sources dominate the discipline of ancient numismatics as presented by Google" (Heath 2010, pg. 41). Nonetheless, this subsection will focus on nonprofit organizations and academic digital projects in numismatics and outline some issues that will need to be addressed to create a digital infrastructure for this discipline.

Numismatics Databases

One of the largest organizations dedicated to the field of numismatics is the American Numismatic Society (ANS). [343] The ANS has perhaps the largest numismatics database available online (over 800,000 items) and provides access to a searchable database of "coins, medals, banknotes and other numismatic items." This database includes extensive images of coins from the ancient world (Hellenistic Greece and the Roman Republican Period in particular). [344] A variety of searching options are available including by denomination, object type, standard reference, and material as well as keyword searching of various fields (artist, mint, obverse type, obverse legend, reverse type, reverse legend, person, region). Many object records include digital thumbnail images, descriptive information such as object type, material, weight, denomination, region, person illustrated on the coin, and also provide a stable URLs for linking. [345]

---

[340] Other research has also explored the creation of annotation technologies for digital manuscript collections and the ability to share them, see for example (Doumat et al. 2008), which examined storing user annotations in a collaborative workspace so that they could be used in a recommender system for other manuscript users.

[341] http://wordnetweb.princeton.edu/perl/webwn?s=numismatics

[342] For one example of an excellent website created by an enthusiast see http://www.snible.org/coins/, and in particular the "Digital Historia Numorum: A Manual of Greek Numismatics" (http://www.snible.org/coins/hn/), a typed in version of the 1911 edition of the "Historia Numorum" by Barclay Head. Ed Snible created the HTML edition with help from some volunteers and also individually scanned the photographs in this collection. Another large website created by an individual enthusiast is "Magna Graecia Coins" http://www.magnagraecia.nl/coins/index.html.

[343] http://numismatics.org/

[344] http://data.numismatics.org/cgi-bin/objsearch

[345] See for example, http://data.numismatics.org/cgi-bin/showobj?accnum=1941.131.932

Another significant online collection with a focus on Roman coins is Roman Provincial Coinage Online (RPC Online). [346] This website has been created by the University of Oxford and has been funded by the AHRC. While this current project is confined to coins from the Antonine period (AD 138-192), one goal of the RPC series project is to produce both a "standard typology of the provincial coinage of the Roman Empire from its beginning in 44 B.C. to its end in AD 296/7" and a model for putting more collections online. The current database that is available is based on 10 collections and includes information on "13,729 coin types based on 46,725 specimens (9,061 of which have images)."

RPC Online provides both a quick search of the whole collection and three specialized types of searches: 1) an *identification* search ("to identify a coin or find a standard reference") where the user can search by city, obverse or reverse design, reverse inscription (includes a Greek keyboard), metal and diameter). The user can also select from a list of cities or features on the obverse or reverse design of the coin. 2) an *iconographic* search that examines the type of imagery used on coins, the user must first choose a design group (animals, architecture, deities, games, heroes, imperial family, object), then choose from a list of relevant options and finally choose to search all provinces or an individual one (e.g. animals—chimera—Achaea); [347] 3) an *advanced* search including mint location, date, magistrate, design & inscriptions, physical characteristics and coin reference. Rather than records of *individual* coins, this database contains records for individual *coin types* with both thumbnail and high-resolution images and a full description. Geographic access is also provided to the collection through a Flash map [348] where a user can either browse a map and choose a city or pick from a list of cities where selecting a city will take them to a list of matching coin types. This database provides numerous means of accessing its collections, and is also unique in that it supported searching of Greek inscriptions found on coins.

The Sylloge Nummorum Graecorum (SNG) [349] is one of the larger Greek numismatics databases available online. This website is a British Academy Research Project and has the major purpose of publishing "illustrated catalogues of Greek coins in public and private collections in the British Isles." SNG has retained a traditional broad definition of Greek "to include the coins produced by all ancient civilisations of the Mediterranean and neighbouring regions except Rome, though it does include the Roman Provincial series often known as 'Greek Imperials'." While the SNG had traditionally focused on print publication it has increasingly utilized electronic publication, and it thus developed a relational database for this website that includes 25,000 coins from the SNG volumes. This database can be searched by a variety of fields including collection, state, mint, material, ruler, period, denomination, hoard, obverse or reverse coin description, among others. Records for individual coins include thumbnail images and full descriptive information. [350]

Another large online collection is the Numismatische Bilddatenbank Eichstätt [351] that provides various means of access to a virtual library of coins from several German universities and museums but particularly from the Catholic University of Eichstatt-Ingolstadt. While the interface is only available in German, this database includes over 5600 objects and there are various ways of accessing the coins. A user can conduct a full text search of all the database fields, browse a list of all words found in the coin legends, or choose a controlled keyword from the thesaurus. In addition, a number of indices to the collection can be browsed including personal names, mint locations, dates, and collection. There are also a variety of ways to browse the entire collection including a short list with images, a standard list of images and descriptions, a picture gallery or a simple list of pictures. Individual coin records include high -resolution images, the name of the person on the coin, the date of the coin, and the collection it is from.

Individual universities and colleges often hold small numismatic collections as well. [352] The University of Virginia Art Museum has recently digitized a collection of nearly 600 Greek and Roman coins. [353] Interestingly,

[346] http://rpc.ashmus.ox.ac.uk/
[347] http://rpc.ashmus.ox.ac.uk/search/icono/?provinces=sel&province-1=Achaea&stype=icono&design_group=4&design-0=114&step=3&next=Finish
[348] http://rpc.ashmus.ox.ac.uk/maps/flash/
[349] http://www.sylloge-nummorum-graecorum.org/
[350] http://www.s110120695.websitehome.co.uk/SNG/sng_reply2a.php?verb=SNGuk_0300_3363
[351] http://www.ifaust.de/nbe/
[352] For example, see the Princeton University Numismatic Collection (http://www.princeton.edu/~rbsc/department/numismatics/).
[353] http://coins.lib.virginia.edu/

this entire collection was described using Encoded Archival Description (EAD), [354] a standard first developed for the description of archival finding aids. The project extended the EAD with several specific adaptations for coins such as to describe physical attributes like iconography. According to the website, this project seems to be the first that has applied EAD to numismatics. They found EAD to be useful for not only did it allow them to describe the physical attributes of each coin, but also they were also to encode "administrative history, essays, and index terms" in XML and thus create sophisticated metadata for searching and browsing. Further technical and metadata details of their implementation have been explained in (Gruber 2009). Gruber explained that one important step in creating EAD descriptions of the object was that subject specialists *normalized* coin legends and personal and place names in order to support standardized searching of proper Latin names and abbreviations (e.g. a search for coins "minted in Greek Byzantium or Roman Constantinople" could thus be accomplished by searching for Istanbul). This encoding of personal names, geographic names, and deities among others was essential for "establishing authority lists for faceted browsing and normalization" and thus made for more sophisticated textual searching. Gruber also reported that they used Apache Solr, [355] an "open source search index based on the Lucene Java library" that includes a number of useful features such as hit highlighting and faceting.

The database for the University of Virginia Art Museum Numismatic Collection can be searched or browsed. A basic search includes the ability to search multiple terms in a variety of fields (keyword, century, collection, deity, denomination, geographical location, iconography, etc.) while an advanced search makes use of Lucene query syntax. This collection also offers a faceted browsing interface where coins can be browsed by different categories (city, collection, deity, material, name, etc.). Records for each coin include multiple images (including high-resolution ones), descriptive information, archival information, bibliography, and a list of index terms such as personal names and subjects that have been linked to entries in Wikipedia and other digital classics resources (Livius.org, Theoi.com). Individual records also have permanent URL's that are easy to cite and a social bookmarking feature is also available for each record. [356] One useful feature is that this entire record can also be printed. Another unique feature of this collection is the ability to compare two coins.

Several smaller individual numismatic collections have also been created as parts of museum exhibitions or educational resources. For example, "Bearers of Meaning: The Ottilia Buerger Collection of Ancient and Byzantine Coins" [357] includes a series of thematic essays and a catalogue of almost 150 coins. Each catalog entry includes images, descriptive information, provenance, bibliography, and a descriptive entry. [358] While the "Bearers of Meaning" website was designed to accompany an educational exhibit, other museums have provided online access to their entire numismatic databases. The Bruce Brace Coin Collection [359] at McMaster University Museum of Art includes 272 Roman coins, a series of thematic tours, and a timeline. The collection can only be searched by a general search box (with no limits). Records for individual coins include images, timeline year, condition, location, provenance, a description and additional notes, and references to numismatic catalogues. One unique feature of this collection is the ability to zoom in on individual images at great detail using a tool called Zoomify.

Another online exhibit is the "Coinage of Ephesus" at Macquarie University in Sydney, University. [360] This website includes a lengthy descriptive essay about the collection that is divided into chapters with links to the individual coins and an interactive gallery of coins (that requires Flash to view). Clicking on a coin brings up images of the coin along with basic descriptive information. There is no way, however, to search the collection of coins. Finally, another useful resource is the "Virtual Catalog of Roman Coins" (VCRC) [361] a website that has is maintained by Robert W. Cape, Jr., Associate Professor of Classics, Austin College, and is "devoted to

[354] http://www.loc.gov/ead/
[355] http://lucene.apache.org/solr/
[356] For example, http://coins.lib.virginia.edu/display-uva?id=n1990_18_3
[357] http://www.lawrence.edu/dept/art/buerger/
[358] For example see, http://www.lawrence.edu/dept/art/buerger/catalogue/033.html
[359] http://tapor1.mcmaster.ca/~coins/index.php
[360] http://learn.mq.edu.au/webct/RelativeResourceManager/15043963001/Public Files/index.htm
[361] http://vcrc.austincollege.edu/

helping students and teachers learn more about ancient Roman coins." This website contains coin images and descriptions from the early Roman Republic through the end of the 4th century A.D. The VCRC can be searched by either a general keyword search or by coin issuer, obverse or reverse description, inscription and contributor. Since the VCRC was designed as an educational resource it also includes a list of student projects and teaching resources.

Some numismatic research databases that once had individual websites have been archived by the Archaeology Data Services. One example is "Analysis of Roman Silver Coins: Augustus to Nero (27 B.C. – AD 69),"[362] a project conducted in 2005 by Matthew Ponting and Kevin Butcher at the University of Liverpool. The research database includes numismatic descriptions of coins and pictures, and can be queried by denomination, mint, emperor, hoard, or donor. This website illustrates the importance of access to digital preservation services for individual faculty research projects once they are completed.

Numismatic Data Integration and Digital Publication

As indicated by this brief overview there are numerous numismatics databases many of which have collections of overlapping time periods and geography but all of which provide varying levels of access through different types of database interfaces and utilize various schemas often with an extensive number of different fields/elements that describe the same data items in different databases. The challenges of integrating such collections are numerous and some of them have been explored by D'Andrea and Niccolucci (2008). These authors examined data harmonization efforts using the CIDOC-CRM ontology and described initial efforts to map three different numismatics databases to the CIDOC-CRM and to develop a "general numismatic reference model." Similarly the CLAROS project has also utilized CIDOC-CRM to provide federated searching to various classical art databases.

This lack of a common standard schema for numismatic databases is not surprising as there is a similar lack of standards for both the cataloguing and analysis of coins within printed publications in this field according to a recent article by Kris Lockyear (Lockyear 2007). In his overview of the recording and analysis of Roman coins in Britain, Lockyear also criticized the English Heritage guidelines that had recently been released for describing coins. One of the major problems, Lockyear suggested, was the identification of coins, for without a "well-preserved genuine coin" even extensive patience and using the ten volumes of *Roman Imperial Coinage* did not necessarily provide a scholar with a distinct catalogue number along with a date range, place of manufacture and denomination. Lockyear noted however, that detailed analysis of sites required including not just well identified coins but all the coins found, including those that were poorly preserved. To address this issue, a series of "coin-issue periods" were created so that coins could at least be assigned to a period and summary listings could be created of the coins at a site. There are currently two such schemes used in Britain and conversion between them typically requires a full catalogue of the coins found, Lockyear reported, but unfortunately, many printed publications included only partial catalogues of site finds.

Lockyear explained that the English Heritage guidelines suggested three levels of cataloging: a full catalogue with detailed information, a shorter catalogue of the full one, or a "spreadsheet" that is typically a summary of data by coin periods. The minimum data to be included was coin identification according to standard catalogues, identification code, site code and small find number. Lockyear stated that the Heritage guidelines regarding the weight and diameter of coins were problematic, and in particular, he disliked the instructions on how to record coin legends. While three database schemas were provided by the guidelines, Lockyear considered them to be poorly designed and as not taking full advantage of relational database capabilities. This was unfortunate, Lockyear noted, because a good database design could make it easy to produce catalogues that conformed to almost any format as long as the necessary data had been entered. Good database design would also reduce duplication of effort at data entry (e.g. a table of legends could be used so a legend did not have to be entered for each coin that exhibited it). In addition, Lockyear also argued that since "fuzzy data" is being recorded for many coins, the ability to indicate levels of certainty for some fields would be very important. Another area for

[362] http://ads.ahds.ac.uk/catalogue/archive/coins_lt_2005/index.cfm?CFID=3825887&CFTOKEN=59064527

improvement Lockyear proposed was in the categories of analysis used in numismatics, and to move beyond simply analyzing when coins were produced to examining "coin-use periods."

The greatest benefit of a well-designed database Lockyear reasoned would be the ability to automatically generate summary lists of coins. Additionally, he proposed linking such databases to GIS packages in order to better enable "intra and inter-site analyses," a possibility that he was surprised had not been explored by more numismatists. Lockyear believed three points were essential to move numismatics forward: 1) all coins from excavations should be identified to the extent possible, 2) a standard database schema should be created that could be used by specialists in the field and easily archived with the ADS, and any such schema should not be dependent on a particular piece of software, 3) the analysis of coins should be integrated with stratigraphic and ceramic data. [363] In order to help start bringing such change about, he proposed making available a "user friendly and flexible database application." An even better solution he suggested would be to develop a web-based system that made use of MySQL or another open source system, for this would support the use of common database tables and make data universally available. Lockyear hoped that new finds would be entered into such a system, that archaeologists would be able to download all or any part of this data and analyze it using their favorite tools, and that legacy numismatic data such as site lists and hoards would also begun to be input. [364] Lockyear concluded by reiterating that the analysis of coinage data needs to be integrated with other strands of archaeological information:

> The future lies, I believe, in the integration of stratigraphic, coinage and other evidence such as ceramic data. This is hardly a revolutionary idea, being commonplace in other contexts, but one which we must now pursue with some vigour (Lockyear 2007).

As was seen in other disciplines such as epigraphy, scholars are increasingly looking to the potential of databases and the digital world to reintegrate different data sources that have often been arbitrarily divided by disciplinary structures, in order to provide a more holistic approach to studying the ancient world.

Perhaps the most significant work regarding the challenges of numismatic data integration is being conducted by the Digital Coins Network, [365] which is promoting "the effective use of information technology in the collection, exchange, and publication of numismatic data." This network both identifies existing standards and promotes the development of new standards for the numismatic community. The current major project of the network is the refinement and extension of the Numismatic Database Standard (NUDS), [366] and they are "working to define a standardized set of fields to describe numismatic objects within the context of a column-oriented database." The network also recognizes that there are already thousands of records of numismatics objects in traditional relational databases so they plan to create a set of shared fields [367] that will promote exchange of data and are developing a NUDS testbed.

In addition to the challenges of data integration, the need to support more sophisticated digital publications in numismatics will be another challenges for any developed infrastructure. The ANS has recently announced a digital publications project [368] and according to their announcement, they are "developing an infrastructure for the digital publication of numismatic catalogs, exhibitions, articles, and other materials." The system will take advantage of existing standards and schemas that are already available so text will be encoded using XML (with TEI DTD's and schemas where appropriate). Currently they already have a number of experimental digital publications including a preliminary HTML map of mints, [369] a growing catalog of "Numismatic Literature" that

---

[363] One useful source for the study of Roman ceramics is "Roman Amphorae: A Digital Resource" (http://ads.ahds.ac.uk/catalogue/archive/amphora_ahrb_2005/info_intro.cfm?CFID=3775528&CFTOKEN=68253066) that is available in the ADS archive. A directory of online ceramics and pottery resources (particularly from the Roman period) can be found at Potsherd (http://www.potsherd.uklinux.net/)

[364] Lockyear also mentioned the possibility of integrating the data from one extensive project, the "Portable Antiquities Scheme", "a voluntary scheme to record archaeological objects found by members of the public in England and Wales." All of these finds are available through a central database and it includes thousands of coins. http://www.finds.org.uk/

[365] http://www.digitalcoins.org/

[366] http://digitalcoins.org/index.php/NUDS:Fields

[367] Initial work on the NUDS:Exchange, "an xml schema designed to facilitate the exchange of numismatic information" can be found at http://digitalcoins.org/index.php/NUDS:Exchange_Format

[368] http://www.numismatics.org/DigitalPublications/DigitalPublications

[369] http://numismatics.org/xml/geography.html

continues their annual listing of numismatic titles, and a number of HTML catalogs of exhibits and traditional publications.

The most advanced of the ANS digital publications, however, is Nomisma.org, [370] a joint effort by the ANS, Yale University Art Gallery, and the Paris-Sorbonne University to "provide stable digital representations of numismatic concepts and entities" such as generic concepts like a "coin hoard" or actual hoards listed in published collections such as the Inventory of Greek Coin Hoards (IGCH). Nomisma.org provides stable URIs [371] for the resources it includes and in the spirit of "linked data" [372] defines and presents the information in both a human and machine-readable form. All resources are represented in XML and they also plan to utilize XHTML with embedded RDFa. The hope is that creators of other digital collections will then use these stable URIs to build a web of linked data "that enables faster acquisition and analysis of well-structured numismatic data." For an initial test case, Nomisma.org is developing a digital version of the IGCH, and all of the 2387 hoards have been given stable URIs. The IGCH was chosen because the hoards identified within it have unique identifiers that are well-known to the community and hoards are typically conceived as lists of links to other numismatic entities (the mints of coins and findspots). This presents the opportunity of defining encoding conventions for these entities and for then turning the information into explicit hyperlinks.

There was surprisingly limited research regarding digitization strategies for coins and the development of digital collections in numismatics. One area of growing research, however, is that of automatic coin recognition. Kampel and Zaharieva (2008) have recently described one state-of-the-art approach:

> Fundamental part of a numismatists work is the identification and classification of coins according to standard reference books. The recognition of ancient coins is a highly complex task that requires years of experience in the entire field of numismatics. To date, no optical recognition system for ancient coins has been investigated successfully. In this paper, we present an extension and combination of local image descriptors relevant for ancient coin recognition (Kampel and Zaharieva 2008)

They identified two major processes that must first be differentiated, coin identification where a unique identifier is assigned to a specific coin and coin classification where a coin is assigned to a predefined type. The authors argued that automatic coin identification was an easier task due to the nature of how ancient coins were created. For example the manufacturing process tended to give coins unique shapes (hammering procedures, coin breakages, etc.) These same features that assist in individual coin identification complicate automatic classification, however, for as they noted "the almost arbitrary shape of an ancient coin narrows the amount of appropriate segmentation algorithms" (Kampel and Zaharieva 2008). Additionally, algorithms that performed well on image collections of modern coins did not fare well on medieval ones.

Ultimately, Kampel and Zaharieva (2008) decided to use "texture sensitive point detectors" and conducted initial experiments to determine what local feature descriptors would work best for identifying a given set of interest points in ancient coins. After acquiring a set of images of 350 different coin types from the Fitzwilliam Museum in Cambridge they built a coin recognition workflow. Since they were using images of individual coins they did not need to automatically detect and segment coins found in the images and instead focused on the feature extraction step. The feature extraction process involved two steps, the use of local feature algorithms to extract local image descriptors that could be used in individual coin identification and then the extraction of features that could be used to "to reduce the number of required feature comparisons" by reducing the coins that needed to be extracted from the database. After an initial pre-selection step, they performed descriptor matching by "identifying the first two nearest neighbors in terms of Euclidean distances." The final step involved a verification process. An algorithm called SIFT (Scale Invariant Feature transform) provided the best results in terms of discriminant feature identification, but its biggest drawback was computational time. For future experiments, they planned to expand their evaluation to a larger set of coin images. As more collections of coins become available for online experimentation, it is likely that the accuracy and viability of such approaches will correspondingly increase.

---

[370] http://nomisma.org/
[371] For example, http://nomisma.org/id/igch0262
[372] For more on publishing linked data on the web, see (Bizer et al. 2007).

Palaeography

The discipline of palaeography has received some brief examination in other sections, such as in the creation of a palaeographic knowledge base for Cuneiform and in terms of automatic document recognition for Latin manuscripts.[373] A fairly comprehensive definition has been offered by Moalla et al. 2006:

> The paleography is a complementary discipline of the philology ….The paleography studies the layout of old manuscripts and their evolutions whereas the classic philology studies the content of the texts, the languages and their evolutions. The goals of the palaeographic science are mainly the study of the correct decoding of the old writings and the study of the history of the transmission of the ancient texts. The palaeography is also the study of the writing style, independently from the author personal writing style, which can help to date and/or to transcribe ancient manuscripts (Moalla et al. 2006).

The study of palaeography is thus closely tied to work with many other disciplines, and as Ciula (2009) explains "palaeography cannot proceed without sharing methods, tools and outcomes with co-disciplines such as epigraphy, codicology, philology, textual criticism—to name but a few." As this comment illustrates, while there are many specialized disciplines within classics, they also shared many research methods. This section will therefore explore some recent state-of-the-art work in digital palaeography.

In a recent paper, Peter Stokes (Stokes 2009) has provided an overview of the issues faced in attempting to create a discipline of digital palaeography. He observed that traditional palaeographic studies have their own methodological issues, in particular a lack of established terminology (e.g. for handwriting), a factor that has made those few digital resources that have been created difficult to use and "almost impossible to interconnect." This has not only frustrated scholarly communication but also made creating "databases of scripts" almost impossible. Nonetheless, there are extensive digital corpora now available, and Stokes argued that such corpora could not be analyzed by traditional methods because they can include "hundreds of scribal hands with potentially thousands or tens of thousands of features." Both the creation of new databases and the use of data mining, Stokes asserted, would be necessary to work with such large bodies of material. Yet digital methods he acknowledged had still received little acceptance from scholars in the discipline:

> However, promising as these seem, they have received almost no acceptance and relatively little interest from 'traditional' palaeographers. This is partly because the technology is not yet mature, and perhaps also because the attempts to date have generally involved small projects without the sustained funding or larger interdisciplinary groups that digital humanities often require (Stokes 2009).

In addition to the challenges of relatively new and untested technology, limited funding, and small projects, Stokes also expressed how the use of digital methods is also problematic because it requires understanding of many fields such as computer graphics and probability theory, skill areas that most traditional palaeographers can't be expected to have.

One potential solution Stokes believed was to develop software that presented results in an intelligible manner to palaeographers. Consequently, Stokes is currently working on a software platform for image enhancement called the "Framework for Image Analysis" a "modular and extendible software in Java for the analysis of scribal hands." This software allows users to load images of handwriting and run various automated processes to analyze and generate metrics for scribal hands. This system also includes a module to enhance images before they are processed that can also be run as a stand alone application to try and recover damaged text from manuscripts. One useful feature of this system is that the user can compare various metrics and distances generated by different processes (implemented as plug-ins) on different pieces of writing, and they can also implement their own algorithms and export the results of these processes. As Stokes noted, this type of system, "allows people to compare different techniques in a common framework, producing libraries of scribal hands and plugins as a common and documented basis for palaeographical study" (Stokes 2009). The ability to create results that could be either "reproducible or at least verifiable" was also important, although Stokes believed that issues of documentation and reproducibility were manageable in that software could be designed to record all actions that are performed and save them in a standard format.

---

[373] Some other computer science research in palaeography has focused on the development of automatic handwriting recognition for medieval English documents (Bulacu and Schomaker 2007) and for 18th and 19th century French manuscripts (Eglin et al. 2006).

Thus Stokes highlighted the need for common frameworks for analysis, the use of standards, and reproducible results to build the foundations for digital palaeography. One other valuable point he made was that designers of digital humanities applications needed not just to consider what algorithms to implement but to how present those results in an intelligible manner to non-computer scientists: "Indeed, it is an important question how the results of complex algorithms can best be presented to scholars in the humanities," Stokes concluded, "and it may well be that the plugins should allow both 'computer- friendly' and 'human-friendly' output, with the latter including graphical or even interactive displays"(Stokes 2009).

Recent work by Arianna Ciula (Ciula 2009) has also explored the methodological issues involved in using digital technology to support palaeographical analysis of medieval handwriting. She maintained that digital methods would only assist palaeographers if the complex nature of the cultural artifacts they studied were also considered. In addition, she also argued that the identification of "critical processes within the palaeographic method" was essential before any tools were developed. Digital tools needed to make the steps of scholarly analysis more explicit, Ciula insisted, including "analyses, comparisons, and classifications. Since palaeography is closely related to many other classical disciplines, Ciula also argued that a more integrated digital environment of *tools* and *resources* was necessary:

> Therefore, independently from its more or less limited scope, the more any digital tool or resource—being it a digital facsimile of a manuscript, an application to segment letter forms, a digital edition, or an electronic publication of other kind—can be integrated within an environment where complementary material is also accessible, the more it becomes exponentially useful to the palaeographer (Ciula 2009).

Palaeographers in particular need more visual representations of manuscripts and open access comprehensive collections.

In her own work, Ciula developed a computing application called "System for Palaeographic Inspection"(SPI) for work she conducted as a graduate student. Ciula scanned the leaves of several codices and developed a basic system that included image pre-processing, insertion of images into a relational database, the segmentation of handwriting in images into relevant letters and ligatures, and the automatic generation of letter models. She created extensive documentation regarding her choice of digitization criteria, the refinement and evaluation of segmentation processes, and tuning the parameters for generating letter models. For this she made extensive use of the large body of literature on manuscript digitization and OCR development, but she also underscored that the development of this system required extensive *domain knowledge* as well:

> On the other hand, the interpretative phase based on the analysis of the letter models and their automatic clustering has required insights into a much more established tradition of *doing* palaeography. The comparison of types of letterforms—which is the main objective of analytical palaeography—has not effectively been supported so far by any tool. Therefore, the major challenge was represented by the attempt to integrate and support the palaeographical method within a digital humanities (as defined by McCarty….) research approach (Ciula 2009).

One of the greatest challenges faced by many digital classics practitioners was the difficulty of needing both extensive disciplinary expertise and technical knowledge to do their work.

The tool she developed had a number of technical limitations, Ciula granted, and she commented that various scholarly stages of interpretation such as letter segmentation and model generation were assisted by the tool but not "comprehensively and systematically supported by the tool itself." The most powerful function of the SPI was its ability to "compute graphical features" and this assisted palaeographic analysis by making variations between characters more perceptible to human vision. Ciula nonetheless emphasized that her tool was meant to *assist* scholars and not replace them and this raised the question of how well digital tools could ever model scholarly expertise. "How much of the palaeographical expertise can the tool or its modules incorporate?" Ciula asked, "If the use of the tool itself contributes to define, refine and enrich the underlying method, to what extent can this process be fed back into the tool and make it more sophisticated?" (Ciula 2009). The iterative process of modeling scholarly expertise in a computational manner and creating a tool that can both utilize this initial knowledge and feed new knowledge back into the system has also been explored by the eSAD project in their development of an interpretation support system for papyrologists (Tarte 2010, Olsen et al. 2009).

In sum, Ciula hoped that the functionality of the SPI could be turned into a paleographical module that could then be utilized as a web service as part of larger infrastructure that might be used in the "creation and annotation of digital editions." Other necessities Ciula proposed for a larger digital research infrastructure included the need for documentation and transparency, the use of standards, systems that were extensible and interoperable, and more funding for interdisciplinary research questions.

Papyrology

As described by Bauer et al. (2008) papyrology "focuses on the study of ancient literature, correspondence, legal archives, etc. as preserved in papyri." Both individual papyri collections and massive integrated databases of papyri can be found online. In fact, digital papyrology is considered to be a relatively "mature" digital discipline. [374] "Repositories for certain primary sources, such as papyri, are already playing an important role in ensuring access to digital surrogates of artifacts," Harley et al. (2010) observed, "In the study of written evidence, these databases of annotated primary sources could also play an important role as digital critical editions." As was seen previously in the discussion of inscriptions the amount of scholarly editing and interpretation that often goes into the production of online images and transcriptions of papyri should be considered akin to the act of creating a critical scholarly edition.

The discipline of papyrology first began to take shape in the late 1880s and 1890s according to an article by Ann E. Hanson, and the importance of these sources for the study of the ancient world was quickly recognized:

> A direct and immediate contact with the ancient Mediterranean was being established, as texts, unexamined since antiquity, were being made available to the modern world. To be sure, this writing paper of the ancients had been used not only for elegant rolls of Greek literature, but also for quite everyday purposes in a variety of languages, accounts, letters, petitions, medicinal recipes. Still, it was the copies of Greek literature which had not survived in the manuscript traditions that were particularly prized in the early days, for these were the more accessible to scholars trained in the authors of the canon (Hanson 2001).

Papyri thus offered access not just to works of lost literature but also to the documents of daily life. In the late 19th and early 20th century, many European and American libraries began to create collections of papyri such as from Tebtunis, [375]Oxyrhynchus, [376] and Herculaneum, [377] collections that are increasingly becoming available online. Despite some competition for collections, Hanson described how there was a considerable amount of collegiality between various scholars and collectors that was described as the "amicitia papyrologorum" and this collaborative nature of papyrological research continues today.

This section will examine a number of digital papyri projects online [378] and also look at several significant research projects seeking to develop new technologies for the analysis of papyri and to create digital research infrastructures for papyrology.

Digital Papyri Projects

The largest papyri projects to be found online are those projects that serve as aggregators, union catalogues or portals to other papyri collections. The Advanced Papyrological Information System (APIS) [379] is one of the oldest and largest papyri databases online and according to its website "is a collections-based repository hosting information about and images of papyrological materials (e.g. papyri, ostraca, wood tablets, etc) located in collections around the world." This repository contains physical descriptions, extensive bibliographic

---

[374] For example, the 26th International Conference on Papyrology has three separate sessions on "Digital Technology and Tools of the Trade" (http://www.stoa.org/?p=1177).
[375] http://tebtunis.berkeley.edu/
[376] http://www.papyrology.ox.ac.uk/POxy
[377] http://www.herculaneum.ox.ac.uk/papyri.html
[378] This section will focus on some of the larger projects as there are numerous interesting projects such as individual university collections that have been digitized such as Harvard's "Digital Papyri at the Houghton Library" (http://hcl.harvard.edu/libraries/houghton/collections/papyrus/index.html) or the Rylands Papyri digitized by the University of Manchester (http://www.library.manchester.ac.uk/eresources/imagecollections/university/papyrus/#d.en.98702). There are also websites that have been dedicated to individual papyri of particular interest such as "Edwin Smith's Surgical Papyrus" (http://archive.nlm.nih.gov/proj/ttp/flash/smith/smith.html), and research projects that are currently focused on the papyri of an individual author such as Philodemus (http://www.classics.ucla.edu/index.php/philodemus).
[379] http://www.columbia.edu/cu/lweb/projects/digital/apis/index.html

information, digital images, and English translations for many of the texts. In some cases, links are provided to the original language texts. As of March 2010, the APIS "union catalogue" included 28,677 records and 18,670 images from over 20 collections of papyri with the largest being Columbia, Duke, New York University, Princeton, University California-Berkley, the University of Michigan, and Yale. [380]

The APIS includes both published and unpublished material and is hosted by the Columbia University Digital Libraries project. The collection can be searched by keyword across the whole collection or within an individual papyri collection. Individual papyri can also be searched for by publication number, collection number, or APIS number. There are also a number of browsing features including by subject word, documentary or literary type, writing material, and language (including Arabic, Aramaic, Coptic, Demotic, Greek, Latin, Hebrew, Hieratic (Egyptian), Hieroglyphic, Italian, Middle Persian, Parthian, and Syriac). An advanced search offers even more options. Each individual papyri record includes an APIS identifier, title, language, physical description, notes, etc. and digital images where available. Some records also include a link back to the original papyri collection database for fuller information that may be provided there. The APIS is also involved in the larger digital classics research project Integrating Digital Papyrology (IDP).

The potential for projects such as the APIS received early recognition from papyrologists such as Ann Hanson. "This wealth of electronically searchable materials means that more possibilities can be explored at every phase in the process of preparing a papyrus for publication," Hanson asserted, "from finding parallels to assist reading to the contextualization of a papyrus' message back into the circumstances that seemed to have occasioned its writing" (Hanson 2001). She also praised the fact that the APIS was greatly expanding *access* to both papyri and papyrological information, particularly through its links to translations, and that making digital resources available to an audience beyond the academic world was a goal for which all projects should strive.

One of the oldest and largest individual papyri collections online, with federated access also provided by the APIS, is the Duke Data Bank of Documentary Papyri (DDBDP). [381] The DDBDP includes the full text of thousands of Greek and Latin non-literary papyri. Initial online access to this collection was provided through the Perseus Digital Library, [382] and Perseus continues to provide a browseable list of texts where the full text of the papyri is available for viewing online. [383] While these texts are searchable through the various advanced searching features of Perseus, a newer papyrological search engine also provides access to this collection and is available at papyri.info, [384] a site that is part of the larger IDP project. The DDBDP is also a project partner in both the IDP and the Concordia projects.

Similar to the APIS, the Papyrus Portal Project [385] seeks to provides its users with a federated search of all "digitized and electronically catalogued papyrus collections in Germany" and to provide an "unified presentation of the search results with the most important information on the particular papyrus." The Papyrus Portal thus provides integrated access to the digitized holdings of 10 German papyri collections, the largest of which is the Papyrus Project Halle-Jena-Leipzig. [386] The funding for this project was provided by the Deutsche Forschungsgemeinschaft (DFG) and the website was created by the University of Leipzig using the open source software "MyCoRe." [387] The Papyrus Portal database has both a English and German interface and can be searched by inventory number, title, language (Arabic, Aramaic, Demotic, Gothic, Greek, Hebrew, Hieratic, Hieroglyphic, Coptic, Latin, Syriac), text type (documentary, literary, unidentifiable, paraliterary), material, location, date, content and free text. Individual records for each papyrus also include links to their original database so the user can go directly to detailed data without having search individual databases again.

---

[380] Some but not all of these collections also include large online databases that also provide separate access to their collection such as the University of Michigan (PMich- http://www.lib.umich.edu/papyrus-collection) and Berkeley (Tebtunis- http://tebtunis.berkeley.edu/)
[381] For a full history of the DDBDP, see http://idp.atlantides.org/trac/idp/wiki/DDBDP
[382] http://www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:DDBDP
[383] One unique feature of access to the DDBDP through Perseus is that Greek terms are cross-referenced to appropriate lexicon entries in the LSJ.
[384] http://www.papyri.info/navigator/DDBDPsearch
[385] http://www.papyrusportal.de/content/below/start.xml?lang=en
[386] http://papyri.uni-leipzig.de/
[387] http://www.mycore.de/

The most significant collection that is included in the Papyrus Portal Project is Papryus-und Ostrakaprojekt Halle-Jena-Leipzig, a collaborative effort of three universities to digitize and provide access to their papyri collections (both published and unpublished). This growing and well-documented database provides various means of access to the papyri collection in both English and German. While the "general" search includes keyword searching in title, inventory or publication number, text type, collection, place, and district with various date limit options, the full text search includes both a Greek and Arabic keyboard for easier text retrieval. The "complex" search involves sophisticated searching of detailed metadata fields for written objects, texts or documents. Two ways of browsing the collection of papyri and ostraca are also available, the user can either browse an index of written objects, fragments or documents alphabetically by papyri title or they can use a faceted browsing interface again for written objects, texts or documents. Once a user chooses a type of object such as texts, they then choose from a series of metadata categories to create a very specific list of documents (e.g. Written objects – Material – Wood). The record for each object [388] includes thumbnail images (larger images can be viewed using a special viewer that loads in a separate browser window), title, collection, publication number, writing material, size, format, type of text, script, language, date, place, and a link to a full text transcription when available. Both the metadata record for the papyri and the text have static URL's for easy citing and linking.

Another papyrus portal that has recently become available is DVCTVS, [389] a project that aims to ultimately serves a national papyrus portal for Spain. The creation of this project involves four organizations: the Universitat Pompeu Fabra, the Consejo Superior de Investigaciones Científicas, the Abadia de Montserrat and the Companyia de Jesús in Catalonia. Three collections will originally be made available through DVCTVS: the Abadia de Montserrat Collection (consisting of 1500 papyrus from Egypt from the Ptolemaic period until the 10th century A.D. and including literary and documentary texts written in Greek, Coptic, Latin, Arabic and Demotic), the Palau-Ribes collection (with currently 100 published papyri from Egypt between the 8th century B.C. and the 10th century A.D. (with approximately 2000 texts in total) and written in various languages (including Greek, Latin, Coptic, Demotic, Hebrew, Arabic and Syriac), and the Fundacion Pastor Collection (a collection of about 400 papyri from the same time period as the two collections above). Currently papyri are being catalogued, digital images are being added, and previously printed texts are being published as TEI-XML files. A digital catalogue is available that includes multiple keyword searching (using Boolean operators) across a large number of fields (e.g. alphabet, associated MSS, author, book, date, date of finding, edition, findspot, language, published title, etc.). A Greek keyboard can also be used to search the digital catalogue. Each papyri record includes extensive metadata and sometimes includes digital images and a Greek text transcription. The website is available in Spanish, Catalan and English.

One of the largest papyri "portals", albeit with a concentration on collections from Ancient Egypt, is Trismegistos. [390] This project serves as an "interdisciplinary portal of papyrological and epigraphical resources dealing with Egypt and the Nile valley between roughly 800 BC and AD 800" and the large majority of these resources are based at the Katholieke Universiteit Leuven. The core component of this portal is the Trismegistos Texts Database [391] that supports federated searching across the *metadata* (currently 113,940 records) of a series of papyrological and epigraphic databases of related projects. The first group of partner projects updates their information directly into the FileMaker Database that underlies the online XML version of Trismegistos Texts and includes: Hieroglyphic Hieratic Papyri (HHP), [392] Demotic and Abnormal Hieratic Texts (DAHT), [393] Aramaic Texts from Egypt (ATE), [394] TM Magic, [395] and the Leuven Database of Ancient Books (LDAB). [396] Each of these databases also have project websites as part of Trismegistos so their collections can

[388] For example, see http://papyri.uni-leipzig.de/receive/HalPapyri_schrift_00001210
[389] http://dvctvs.upf.edu/lang/en/index.php
[390] http://www.trismegistos.org/
[391] http://www.trismegistos.org/tm/index.php
[392] http://www.trismegistos.org/hhp/index.php
[393] http://www.trismegistos.org/daht/index.php
[394] http://www.trismegistos.org/ate/index.php
[395] http://www.trismegistos.org/magic/index.php
[396] http://www.trismegistos.org/ldab/index.php. Further discussion of the LDAB can be found here.

be searched separately. Metadata also comes from four other major projects that maintain separate databases and update their data in Trismegistos yearly, and this includes the HGV, [397] the Arabic Papyrology Database (APD), [398] the Brussels Coptic Database (BCD), [399] and the Catalogue of Paraliterary Papyri (CPP). [400] Each papyri is given an individual Trismegistos number so that records for it can be easily found across all of the different databases. [401] The Trismegistos portal also provides access to a variety of other major resources. [402]

Of the four separate databases that are federated through Trismegistos, the HGV is the largest, and it is also a participating partner in both the APIS and the IDP projects as well as LaQuAT. First funded in 1988, this database includes extensive *metadata* [403] on almost all Greek and Latin documentary papyri and ostraca from Egypt and nearby areas that have appeared in over 500 print publications. [404] The metadata in the HGV describes the papyri found in many other papyri collections, including many of those found within the APIS and in particular the DDBDP. The current database includes 56,100 records (though each record is not for an individual papyri) since many individual papyri have been published in separate collections and thus there are often several metadata records for the same papyri. The database is based on FileMaker and for those users unfamiliar with this commercial database searching the HGV requires referring to the specific database documentation. The user can also browse an alphabetical list of texts and the record for each papyrus includes details of its publication and links to the full text when available in both the Perseus and papyri.info implementations of the DDBDP.

Another database federated through Trismegistos but that also maintains a separate website for its database is the Arabic Papyrology Database (APD), a project that has been created by the International Society for Arabic Papyrology. The APD allows users to search for Arabic documents on papyrus, parchment and paper from the 7th through 16h century A.D. The website notes that although there are more than150,000 Arabic documents conserved on papyrus and paper, only a small number of these documents have been published and extensively studied. [405] The APD provides access to about 850 (out of 2,000) Arabic texts and is the first electronic compilation of Arabic papyri. Both simple and advanced searching options are available, and the APD supports lemmatized searching of the papyri text and a full search of the metadata. The collection of papyri can also be browsed by name, metadata or references. Each individual papyrus record includes full publication metadata, the full Arabic text (including variant readings and apparatus), a transcription, and relevant lexicon entries for words.

The third database federated through Trismegistos with a significant separate web presence is the Brussel's Coptic Database (BCD), a database of Coptic documentary texts that was started in 2000 and is currently maintained by Alain Delattre of the Centre de Papyrologie et d'Épigraphie Grecque of the Université Libre de Bruxelles. The BCD used the HGV as its initial model and now provides access to about 6700 Coptic texts that have been previously published. The search interface supports multiple fielded keyword searching in French in the following fields: sigla, inventory number, material, origin, date, dialect, content, bibliography, varia, and text ID. Although the website indicates plans to provide access to the full text of documents, currently most of the database is limited to metadata regarding the papyri.

---

[397] http://aquila.papy.uni-heidelberg.de/gvzFM.html
[398] http://orientw.uzh.ch/apd/project.jsp
[399] http://dev.ulb.ac.be/philo/bad/copte/base.php?page=accueil.php
[400] http://cpp.arts.kuleuven.be//
[401] Further discussion of the federated search approach and the use of a unique Trismegistos identifier can be found in (Bagnall 2010) and is also discussed later in this paper.
[402] These include a database of collections of papyrological and epigraphic texts (http://www.trismegistos.org/coll/index.php) created by the Leuven Homepage of Papyrus Collections and the project Multilingualism and Multiculturalism in Graeco-Roman Egypt, a list of papyrus archives in Graeco-Roman Egypt (http://www.trismegistos.org/arch/index.php), the Prosopographia Ptolemaica (http://ldab.arts.kuleuven.be/prosptol/index.html) and a database of place names (http://www.trismegistos.org/geo/index.php)
[403] Another significant database of papyrological metadata is the Mertens-Pack 3 database (http://www2.ulg.ac.be/facphl/services/cedopal/pages/mp3anglais.htm) that provides a catalogue of information and bibliographic details on approximately 6000 Greek and Latin literary papyri.
[404] http://www.rzuser.uni-heidelberg.de/~gv0/Liste_der_Publikationen.html (though a number of journal publications have not been covered)
[405] A full list of published texts from which papyri have been taken can be found here (http://dev.ulb.ac.be/philo/bad/copte/base.php?page=accueil.php)

The final major separate resource federated through Trismegistos is the "Catalogue of Paraliterary Papyri (CPP)" [406] a research project sponsored by the Onderzoeksraad K.U.Leuven and directed by Marc Huys. This electronic catalogue of paraliterary papyri "contains descriptions of Greek papyri and other written materials which, because of their paraliterary character, cannot be found in the standard electronic corpora of literary and documentary papyri, the Thesaurus Linguae Graecae (TLG) and the Duke Data Bank of Documentary Papyri (DDBDP)," making it very difficult for all but the specialist to find them. These paraliterary papyri have typically been published in various editions so the CPP has sought to create a unified collection of these materials. The CPP includes digital versions of full text editions of the paraliterary fragments (in both beta code and Unicode) and all papyri have been encoded in TEI-XML, but are presented online in HTML.

One of the more unique papyrology projects is the Vindolanda Tablets Online, [407] a digital collection that provides access to the online edition of the Vindolanda writing tablets that were "excavated from the Roman fort at Vindolanda in northern England." This website includes searchable editions of Volumes 1 and 2 of the tablets, an introduction, the archaeological and historical context and a reference guide to their use. This website received funding through the Mellon Foundation as part of the "Script, Image, and the Culture of Writing in the Ancient World" program and was created through the collaboration of the Centre for the Study of Ancient Documents (CSAD) [408] and the Academic Computing Development Team [409] at Oxford University. The collection of Vindolanda tablets can be either browsed or searched and the user can search for Latin text specifically or do a general text search within the tablet description or English translation (when available). The tablets database can also be browsed by different categories including "highlights," tablet number, subject, category, type, people, places, military terms, and archaeological context. The record for each tablet includes an image that can be viewed in great detail through a special viewer and also provides a transcription (with notes and commentary by line). Another particular useful feature is the "print friendly tablet display" that provides a printable page of all of the tablet information. [410] Each individual tablet has been encoded as a separate EpiDoc XML file and the custom EpiDoc DTD, the Vindolanda XSL stylesheet and the corpus of inscriptions can all be downloaded. [411]

As has been illustrated above, not only can representations of individual papyri often be found in various databases but also the amount of information available (metadata, images, transcriptions, translations, etc.) can vary significantly between databases. [412] The challenges of integrating such collections are numerous and these issues will subsequently by explored in depth in the next section.

Integrating Digital Collections of Papyri and Digital Infrastructure

Papyrology is such an important discipline with the field of classics that a number of the major digital classics research projects have a papyrology component including Concordia, eAQUA, eSAD, IDP, and LaQuAT. While Concordia and LaQuAT seek to integrate papyri collections with other digital classical resources such as epigraphical databases into larger "virtual" collections that can be simultaneously searched, both eAQUA and eSAD are developing technologies to assist papyrologists in the interpretation of their ancient texts.

Focused exclusively on papyri collections, the IDP project, [413] which is a joint effort of the oldest digital resource in papyrology the DDBDP, the HGV, and the APIS, is working to create a single interface to these

[406] http://cpp.arts.kuleuven.be/index.php
[407] http://vindolanda.csad.ox.ac.uk/
[408] http://www.csad.ox.ac.uk/
[409] http://www.oucs.ox.ac.uk/acdt/
[410] For example, see http://vindolanda.csad.ox.ac.uk/4DLink2/4DACTION/WebRequestQuery?searchTerm=128&searchField=printFriendly&searchType=number&printImage=yes&printCommentary=yes&printNotes=yes&printLatin=yes&printEnglish=yes
[411] http://vindolanda.csad.ox.ac.uk/tablets/TVdigital-1.shtml
[412] For example records for the papyri in one important collection such as the Oxyrhynchus papyri (POxy) can be found in APIS, HGV, and Trismegistos and the full text of many of the documents can be found in the DDBDP. POxy also has a descriptive website (http://www.papyrology.ox.ac.uk/POxy/) and an online database http://163.1.169.40/cgi-bin/library?site=localhost&a=p&p=about&c=POxy&ct=0&l=en&w=utf-8. This database allows the papyri to be searched as well as browsed by papyri number, author, title, genre, data, or volume number. Individual papyri records include a link to a papyri image that is password protected.
[413] http://idp.atlantides.org/trac/idp/wiki/

three collections, a project that has largely been realized through the creation of the Papyrological Navigator (PN). [414] Active research work on improving the PN is ongoing as illustrated by a recent blog post by Hugh Cayless (Cayless 2010c). One particular component of the PN that he has recently improved is a service that provides "lookup of identifiers" of papyri in one collection and "correlates them with related records in other collections." While this service was originally based on a Lucene based numbers server, he is working to replace it with a RDF triple store. One particular challenge is that of data integration and the difficulties of modeling the relationships between the same items in different databases. The complicated nature of these relationships includes several dimensions such as different levels of hierarchy in database structures and various FRBR type relationships (e.g. the ancient document is the *work* but then it has various *expressions* in different printed editions (including translations) and each of those editions has various manifestations (HTML, EpiDoc transcriptions, etc.)). In addition, while the relationships between papyrological items and their *metadata* in different databases can sometimes have a 1:1 relationship (such as is usually the case between the DDBDP and the HGV) there can also be overlap such as between the APIS and the other two databases. Each individual database also has complicated *internal* relationships for while the HGV utilizes the idea of a "principal edition" and chooses a single *canonical* publication of a papyrus; they also include other earlier publications of the same papyrus in their metadata. The DDBDP also follows the same basic idea but instead creates a new record that links to stub records for the older editions of papyri.

In order to better represent the complexity of these relationships, Cayless graphed them in Mulgara [415] (a scalable RDF database that is based on Java), so that he could use SPARQL queries to fetch data and then map these to easily retrievable and citable URLs that follow a standard pattern. Results from SPARQL queries will also be made available as Notation3 [416] and JSON formats in order to create both human readable and usable machine interfaces to the data available through the PN. Cayless also reported that he was looking into using the DC TERMS vocabulary as well as other relevant ontologies such as the FRBR vocabulary. [417] Ultimately Cayless also hoped to link the bibliography in individual papyrus records up to Zotero [418] and to ancient places names in Pleiades. "It all works well with my design philosophy for papyri.info," Cayless concluded, "which is that it should consist of data (in the form of EpiDoc source files and representations of those files), retrievable via sensible URLs, with modular services surrounding the data to make it discoverable and usable."

An in-depth explanation of the entire IDP project has also been offered in a recent article by Roger Bagnall. As explained by Bagnall, the goals of the IDP have changed rather significantly since it was first conceptualized in 1992 in two specific ways:

> One is toward openness; the other is toward dynamism. These are linked. We no longer see IDP as representing at any given moment a synthesis of fixed data sources directed by a central management; rather, we see it as a constantly changing set of fully open data sources governed by the scholarly community and maintained by all active scholars who care to participate. One might go so far as to say that we see this nexus of papyrological resources as ceasing to be "projects" and turning instead into a community (Bagnall 2010).

The IDP like many other digital classics and humanities projects is shifting away from the idea of creating static project-based and centrally controlled digital silos to dynamic community maintained resources where both the data and participation are open to all those scholars who wish to participate. Bagnall also argued that this shift included re-conceptualizing what it means to be an editor and that the distinction that was once made between editing texts and creating textual banks should be abandoned.

As part of this new level of openness, the IDP plans to expose both their data and the code that was used to create the system. This means that if other scholars wish to create a new interface to the data or reuse it in various ways they can do so. The IDP also hopes in the future to include more sources of data and Bagnall lists at least two projects that are reusing their code. The data in the IDP is utilizing the EpiDoc encoding standard, which although created for inscriptions, has been increasingly used for recording papyri and coins. As EpiDoc

---

[414] http://www.papyri.info
[415] http://www.mulgara.org/
[416] Notation3 or N3 is a "shorthand non-XML serialization of Resource Description Framework models, designed with human-readability in mind." http://en.wikipedia.org/wiki/Notation3
[417] http://vocab.org/frbr/core.html
[418] http://www.zotero.org/

uses standard TEI elements, this means that new types of search interfaces can be created that "will interrogate a range of ancient sources of different types." In fact, the Concordia project has begun to create a prototype for this kind of integration between papyrology and epigraphy and also connects the documents to the Pleiades database.

In this second phase of the IDP, they have created an online editing system that will allow authorized participants to enter texts into the DDBDP and metadata into the HGV and APIS. Furthermore, this system will support the creation of editions that will only become publicly visible when the editor chooses to do so. Where previously texts that were published in printed editions had to be retyped into a database, the IDP supports a new form of dynamic publication that is controlled by the individual editors and a larger editorial board. In a manner very similar to the Pleiades model, any user can contribute variant readings, corrections, new texts, translations, or metadata, with all suggestions having to be approved by the editorial board. The editing system records every step of this process from proposal through vetting to a final status as accepted or rejected, and a prose justification must be given at each step. Accepted proposals can also be kept as limited access if the creator desires. One strength of this particular model is that rejected proposals are not deleted forever, and are instead retained in the digital record, in case new data or better arguments appear to support them. Additionally all accepted proposals are attributed to their contributor so that proper scholarly credit can be given to them.

Despite assurances, however, that proper credit would be given, Bagnall noted that many contributors were worried about the *visibility* of their work being diminished as their data became "absorbed" in a larger system. To address this issue, Bagnall reported that the Trismegistos project (that provides access to a number of databases created both internally and externally such as the HGV) adds a unique item identifier to the records for an individual item in each database. This allows the federated system to find all hits for individual items while still keeping databases entirely separate; the user has to move between databases to look at the relevant information. Bagnall thus explains that: "highly distinct branding is central to their approach." While Trismegistos has built in links to both the DDBDP and APIS, users must go to the individual databases from Trismegistos. In addition, no data are exposed for web services. Nonetheless, conversations are apparently underway to try and more closely integrate Trismegistos with the PN, and Bagnall acknowledged that originally the HGV maintained a similarly more closed model before deciding to let the new online editing environment operate on their metadata as well. Thus at the end of IDP both HGV and DDBDP will issue archival XML data under CC licenses.

While Bagnall considered concerns about branding and attribution to be legitimate ones due to issues of funding, credit and tenure, he also thought that the risks of creating closed collections were more onerous:

> But keeping data in silos accessible only through one's own interface has risks too, and in my view they are greater—the risk that search engines will ignore you and you will therefore reach a much smaller audience. Our purpose in existing is education; the more we shut out potential users who will come at the world through Google or similar engines, the fewer people we will educate. That to me is an unacceptable cost of preserving the high relief of the branded silo. Moreover, these resources will never reach their full value to users without extensive interlinkage, interoperation and openness to remixing by users (Bagnall 2010).

In addition to concerns about branding, the other major fear of most scholars was the problem of quality control. Bagnall convincingly argued, however, that the editorial structure of Pleiades and IDP in some ways offered stronger quality control measures in that incorrect or inappropriate information could be removed far more quickly than from a printed text, and that the open system allows the community to alert the editors to mistakes they may have missed. "These systems are not weaker on quality control," Bagnall offered, "but stronger, inasmuch as they leverage both traditional peer review and newer community-based 'crowd-sourcing models." New peer review models such as the ones developed by this project are essential in any digital infrastructure that hopes to gain buy in from a large number of scholars. [419]

Another major point of contention for many scholars Bagnall listed and one not that easily addressed is the issue of personal control. Many scholars are possessive of projects that they have created, and this idea of personal

[419] The importance of new peer review models that make use of the Internet to solicit a broader range of opinions on scholarly material was the subject of a recent *New York Times* article (Cohen 2010).

ownership of objects and data was also strongly illustrated in archaeology (Harley et al. 2010). While personal investment has its merits, Bagnall also put forward that "Control is the enemy of sustainability; it reduces other people's incentive to invest in something." Regarding the experience of IDP, Bagnall worried that too much discussion centered upon *revenue* rather than *expense*, and strongly doubted that there was any "viable earned-income option for papyrology." While the IDP briefly considered direct subscription charges and also pondered creating an endowment, they ultimately abandoned both ideas.

Although they considered it likely that they could raise some money, the IDP was also both uncertain how much money would be needed and what they wanted to "fund in perpetuity." They increasingly realized that if neither the APIS nor the DDBDP were "defensible silos," than neither was the discipline of papyrology. Far more essential than preserving individual projects, Bagnall reasoned, was the necessity of developing a shared set of data structures and tools to exploit various types of ancient evidence. As previously noted by Roueché in her overview of digital epigraphy, many of the disciplinary divisions so entrenched today are in many ways as Bagnall eloquently expresses "arbitrary divisions of a seamless spectrum of written expression" that includes numerous sources. Sustainability for the IDP, Bagnall proposed "will come in the first instance from sharing in an organizational and technological infrastructure maintained to serve a much wider range of resources for the ancient world (and perhaps not necessarily limited to antiquity, either)" (Bagnall 2010). While technological infrastructure will be one cost, the other major costs will be content creation and maintenance. The way forward for papryology, Bagnall concluded, was to go beyond its limits as a discipline and in particular its separateness. Indeed, as illustrated by this review many scholars have commented on the fact that the digital environment has rather unexpectedly provided new opportunities to both transcend disciplinary boundaries and promote a more integrated view of the ancient world.

EpiDoc, Digital Papyrology and Reusing Digital Resources

As illustrated above, EpiDoc is being used by the IDP project to integrate several different papyrology projects. The use of this standard has also allowed researchers to explore new questions using the Vindolanda tablets. Recent work by eSAD [420] (eScience and Ancient Documents), an ongoing project between the e-Research Centre and Centre for the Study of Ancient Documents and Engineering Science at University of Oxford, has examined the various ways in which the highly granular encoding of the Vindolanda tablets can "be used to create a reusable word and character corpus for a networked e-Science system and other e-Science applications" (Roued 2009). The eSAD project has two major goals: first, to develop e-Science tools that aid in interpreting damaged texts, and second, to develop new image analysis algorithms that can be used with digitized images of ancient texts. In terms of Vindolanda, Roued investigated how the encoded EpiDoc XML of the tablets could be used to create a knowledge base of Latin words for an Interpretation Support System (ISS) that would assist users in reading other ancient documents.

The Vindolanda project had decided to use EpiDoc in order to support at least a minimal level of semantic encoding for the tablets. Roued described how even with standard conventions such as Leiden, not all of the conventions were applied evenly, as some scholars used "underdots" to indicate partially preserved characters while others used it to demonstrate doubtful characters. The use of EpiDoc consequently addressed these types of issues with Leiden encoding as it was commonly practiced:

> This example illustrates the primary advantage of encoding the editions in XML. If editors wish to differ between uncertain characters and broken characters they can encode them with different tags. They can then transform both tags into under-dots if they still wish to present both instances as such or they can decide to visualize one instance, underlined and the other under-dotted to distinguish between them (Roued 2009).

Thus the use of EpiDoc allows for different scholarly opinions to be encoded in the same XML file since content markup (EpiDoc XML) and presentation (separate XSLT sheets) are separated. Roued also supported the argument of Roueché (2009) that EpiDoc encoding is not a "substantial conceptual leap" from Leiden encoding.

[420] http://esad.classics.ox.ac.uk

While the first two Vindolanda tablet publications were encoded using EpiDoc, Roued observed that the level of encoding was not very granular and the website as it was designed was not well set up to exploit the encoding. She also made the important point that the level of encoding a project chooses typically depends both on the technology chosen and the anticipated future use of the data. For the next series of Vindolanda tablets, Roued explained that the project decided to pursue an even more granular level of encoding including words and terms in the transcription, which has supported both an interactive search functionality and added greater value to the encoding as a knowledge base. To begin with, the project encoded the tablets in greater detail regarding Leiden:

> Encoding instances of uncertainty, added characters and abbreviations enables us to extract these instances from their respective texts and analyze them. We can, for example, count how many characters in the text or texts are deemed to be uncertain. Similarly, we can look at the type of characters that are most likely to be supplied. These illustrate the many new possibilities for analyzing the reading of ancient document (Roued 2009).

In addition to more extensive encoding of the texts in EpiDoc, the eSAD project also decided to manually perform a certain amount of "contextual encoding" of words, people, place names, dates, and military terms, or basically all of the items found in the indices. For words, the index contained a list of lemmas with references to places in the text where corresponding words occurred, and encoding this data allowed them to extract information such as the number of times a lemma occurred in the text. During the encoding of the indices, the project also discovered that there were numerous errors that needed to be corrected. All of this encoding has been performed in order to support new advanced searching features with a new launch of the website as Vindolanda Tablets Online 2.0 in 2010. In particular, they have developed an interactive search feature using AJAX, [421] LiveSearch, JavaScript and PHP [422] that gives the user feedback while typing in a search term. In the case of Vindolanda, it will give users a list of all words, terms, names and dates that contain their search pattern.

The XML document created for each inscription text contains all of its relevant bibliographic information and textual encoding and Roued explained that this necessitated developing methods that could extract *relevant* information only depending upon the need. The project thus decided to build RESTful web services using the ZEND framework [423] and PHP. The Vindolanda web services receive URL's with certain parameters and return answers as XML, and this allows other projects to utilize these encoded XML files and in particular the knowledge base of Latin words. In particular this web service is being used in their related project that seeks to develop an ISS for readers of ancient documents. The prototype [424] includes a word search that "takes the partially interpreted characters of a word and attaches them to the web service URL as a pattern, thus receiving suggestions for the word using the Vindolanda tablets as a knowledge base" (Roued 2009).

The research work by eSAD with the Vindolanda tablets demonstrates how the use of standard encoding such as EpiDoc can support new research such as the development of knowledge bases from encoded texts and also illustrates the potential of providing access to such knowledge resources for other digital humanities projects through web services.

Collaborative Workspaces, Image Analysis and Reading Support Systems

The unique nature of working with papyri has made it a fairly collaborative discipline, in contrast to some of the other sub-disciplines of classics. Hanson earlier described the "amicitia papyrologorum" of the nineteenth and twentieth centuries and this trend it appears continues today. "Individual specialists, particularly in the study of cultural artifacts or documentary remains, work with collections of artifacts, texts, artworks, and architecture that may span several excavation sites," Harley et al. (2010) explained, "As a result, scholars can be highly collaborative," "in how they locate and work with these materials in order to extract as much information and detail as possible."

---

[421] AJAX, short for "Asynchronous JavaScript and XML" and is a technique "for creating fast and dynamic web pages." http://www.w3schools.com/ajax/ajax_intro.asp;

[422] PHP, stands for "Hypertext Processor" and is a server side scripting language, http://www.w3schools.com/php/php_intro.asp

[423] http://framework.zend.com/

[424] This prototype is discussed in the next section of this paper.

In their review of archaeology, Harley et al. also observed that some scholars desired Web-based workspaces that could be shared when working on documentary remains, and quoted one at length:

> It would be nice to be able to have a more convenient way of looking at images all at the same time and manipulating them. We can now do that with text pretty easily, but let's say a few of us are working on an edition of a papyrus and we want to discuss some particular feature in a high-resolution image of it. The only thing we can really do very easily at that point is to all look at the same webpage or to pass the image around by email and give verbal cues to navigate to a particular point (Harley et al. 2010, pg. 111).

Some initial work towards providing such an environment has been conducted by the VRE-SDM project [425] and this work largely continues under the eSAD project. Tarte et al. (2009) have observed that greater accessibility to legible images and a collaborative working environment are both important components of any potential VRE:

> For Classical historians, adding to the legibility challenge, access to the document is often limited. Collaborative work on the documents is one factor that facilitates their deciphering, transcription and interpretation. The Virtual Research Environment for the Study of Documents and Manuscripts pilot software (VRE-SDM)….was developed to promote non-colocated work between documentary scholars, by providing them with a web-based interface allowing them to visualize and annotate documents in a digitized form, share annotations, exchange opinions and access external knowledge bases (Tarte et al. 2009)

The development of this prototype and the in-depth examination of the working methodologies of scholars that work with such damaged texts through a video-based ethnographic study has been described in Bowman et al. (2010) and de la Flor et al. (2010). de la Flor et al. reported that the use of image processing techniques is not particularly new in either epigraphy [426] or papyrology, but few technologies have been based off of detailed studies of the actual working practices of classicists with digital images. While the project was developed with papyrologists and epigraphers in mind, Bowman et al. (2010) also hoped that the VRE-SDM might prove useful to any scholars that worked with manuscripts and thus they attempted to develop a tool that could be generalized for various disciplines.

In this particular case, however, de la Flor et al. (2010) had videotaped the collaborative work sessions of expert classicists who were working with the Tolsum Tablet, a wooden tablet dating from the first century A.D. "The aim of the filming," Bowman et al. (2010) explained, "was to discover and document the inherent practices, tools and processes used to decipher ancient texts and to establish ways in which a VRE might emulate, support and advance these practices" (Bowman et al. 2010, pg. 95). The VRE-SDM project wanted to both construct and test their interface so they filmed four meetings between three specialists. The scholars worked with the VRE-SDM prototype and were also able to display images of the tablet on a large screen.

By watching how the scholars examined the tablet and progressed in their interpretation of the text, de la Flor et al. observed a number of significant processes at work including: 1) how the scholars identified shapes and letters in order to figure out words and phrases and how this was an iterative process that could involve major reinterpretation of earlier scholarly hypotheses regarding words 2) how scholars drew off their background knowledge in various languages, ancient history, and palaeographic expertise to analyze not just individual letters or words but the text as a whole. In this particular experiment, the ability to enhance multiple digital images of the text and to work collaboratively led the scholars to reinterpret several letters and this consequently led them to reinterpret the word "bovem" as "dquem." They thus concluded that the Tolsum Tablet was not about the sale of an ox but may have instead been an example of an early loan note. [427]

The VRE-SDM prototype as it currently exists is controlled by either mouse or keyboard and according to de la Flor et al. (2010) provides a collaborative workspace where classicists can select high resolution digital images, manipulate them in different ways, use different algorithms to analyze them, and then view them along with images, texts and annotations. An annotation feature is included that allows classicists to comment on letters, words and phrases and to enter translations of them. In addition, to enable users to select and annotate image

[425] http://bvreh.humanities.ox.ac.uk/VRE-SDM.html
[426] The use of advanced image processing technologies used in epigraphy has been discussed earlier in this paper.
[427] For more on the reinterpretation of the text, see (Bowman et al. 2009) and (Tarte 2010).

regions, the VRE-SDM extended the Annotea vocabulary. [428] Using a portlet, users can create, save and share annotations and Bowman et al (2010), reported that they were hoping to build a system towards shared reading that along with the use of a standard format such as EpiDoc XML would allow users to create digital editions that could be "supported by an audit trail of readings." In addition, in order to support integration with other projects, all annotations and metadata are represented as RDF and stored in a Jena triplestore.

While their current funding has only allowed for a pilot implementation, the VRE-SDM project is also considering developing some new functionalities including the creation of "hypothesis folders" where researchers could track translations proposed by colleagues for different texts (de la Flor et al. 2010). Such a feature would be used to allow scholars to associate specific images of a manuscript with translations or other assertions made about parts of that manuscript. They also seek to extend the currently existing annotation tool with an ability to annotate parts of images so that scholars could store the reasons they used to propose translations of letters, words or phrases. This need to annotate images at the level of individual words has also been reported by Cayless (2008) and Porter et al. (2009). One drawback to the prototype was also reported however, and that was that the magnification of images made it difficult to "point at a mark that differs considerably in scale from the original" (de la Flor et al. 2010). Nonetheless, analysis of classicists' actual use of their prototype confirmed their hypotheses that scholars need to be able to select different versions of the same image and to be able to "browse, search and compare images."

In addition to allowing classicists to perform traditional tasks more efficiently, de la Flor et al. also proposed that porting the model of their VRE-SDM to a larger infrastructure might support further re-analysis of ancient documents as seen in their case study on a far larger scale:

> The VRE might be able to provide technological support when such re-interpretations are made. For example, by systematically annotating texts with tentative or firmer analyses of readings it may be possible to provide a way of tracking the consequences of re-interpretation for other similar texts. Currently, classicists draw on their expertise to consider a text, shifting back and forth from analyses of letters to analyses of words, lines of text and eventually to the tablet as a whole. Paleographers tend to use their own drawings of letter forms, developed through their own research. An e- Infrastructure might not only be able to distribute these resources between scholars, but it might also provide the means to communicate, explain and defend justifications, assertions and claims about a text (de la Flor 2010).

The ability of an infrastructure to distribute specialized knowledge resources between scholars and to record and support varying scholarly interpretations of a text are both important components of developing a larger cyberinfrastructure for classics.

The eSAD project has complemented this work on image analysis systems within VREs in their efforts to develop an interpretation support system (ISS) for papyrologists, epigraphers and palaeographers that will assist them as they decipher ancient documents. Similar to the work reported by de la Flor (2010), Olsen et al. (2009), Roued-Cunliffe (2010), and Tarte (2010) have examined the work of papyrologists in detail, particularly the processes used in creating interpretations of texts, in order to model these processes with digital methods. They plan to create a system that can aid the analysis of ancient documents by tracking how these documents are interpreted and read. "Such a system will facilitate the process of transcribing texts," argued Olsen et al., "by providing a framework in which experts can record, track, and trace their progress when interpreting documentary material." At the same time, Tarte (2010) also insisted "the aim of the ISS that is being developed is not to automate the interpretation process, but rather to facilitate the digital recording and tracking of the unravelling of that process." In other words, eSAD does not conceive of creating an intelligent system that will automate the work of scholars but instead is designing a tool that will *assist* scholars as they read ancient documents and help them perform difficult tasks. "In this case," Roued-Cunliffe explained, "these tasks would mainly be capturing complicated reasoning processes, searching huge datasets, accessing other scholars' knowledge, and enabling co-operation between scholars working on a single document" (Roued-Cunliffe 2010).

---

[428] http://www.w3.org/2001/Annotea/. As seen throughout this paper, many different types of annotations (editorial commentary, image annotations, linguistic annotations) are used by various digital classics project. The importance of being able to share different types of annotations both within and across disciplines has led to the creation of the Open Annotation Collaboration (http://www.openannotation.org/), which is currently building a OAC data model and use cases for sharing annotations.

This tool will thus model the tacit knowledge and working processes of papyrologists as well as learn from their behavior in order to expedite both their daily work and make suggestions in future work.

Although the application named DUGA that has been created by the eSAD project is based on decision support system technology (DSS) such as that used by doctors and engineers, they ultimately decided it was an *interpretation* support system they were creating since "experts transcribing ancient documents do not make decisions based on evidence but instead create interpretations of the texts based on their perception" (Olsen et al. 2009). At the same time, one of the key research goals was to explore issues of technology transfer, or to see if the ideas involved in creating a DSS could be transferred to the work of classical scholars (Roued-Cunliffe 2010).

One key idea behind the ISS is that an interpretation is made up of a network of "percepts" that range from low-level (determining a character was created by an incised stroke) to high-level (determining that several characters make up a word). While this network of percepts is implicit in the process of papyrologists, the eSAD project plans to make them explicit in a "human-readable format through a web-based browser application" (Olsen et al. 2009). In the application, the most elementary percepts will be image regions that contain "graphemes" and these images will then be divided into cells "where each cell is expected to contain what is perceived as a character or a space" (Olsen et al. 2009). The division of the image is considered to be a tessellation and documents can be tessellated in different ways. The basic idea is that individual interpretations can be represented as "networks of substantiated percepts" that will then be made explicit through an ontology. "The ontology aims to make the rationale behind the network of percepts visible," Olsen et al. explained, "and thus expose both: (a) some of the cognitive processes involved in damaged texts interpretation; and (b) a set of arguments supporting the tentative interpretation" (Olsen et al. 2009). The final ISS system will use this ontology (that will be formatted in EpiDoc) as a framework to assist scholars in creating transcriptions of texts.

Another step in the modeling process for the ISS was creating image capture and processing algorithms that could embody perceptual processes of papyrologists. As papyrologists often do not have access to the original objects, they frequently work with digital photographs and Tarte (2010) acknowledged that digitizing wooden stylus tablets as "text bearing objects" was not an easy feat. Upon observing papyrologists, they concluded that both manipulations of the images and prior knowledge played important roles in the perception of characters and words. The tablets were visualized using polynomial texture maps and several algorithms were used to detect the text within the images. The algorithms created for minimizing background interference (flattening the grain of the wood) have also been utilized in the VRE-SDM. One of the most complicated (and still ongoing) tasks was developing algorithms to extract the "strokelets" that form characters (including broken ones), as this is the feature "on which the human visual system locks" (Tarte 2010). The final major algorithm developed was a "stroke-completion algorithm" that was created to help facilitate both automatic and scholarly identification of characters. The ISS to be developed will eventually propose potential character readings (utilizing a knowledge base of "digitally identified list of possible readings") but will never force a user to choose one (Tarte 2010).

Many of the insights for both the algorithm development process and the format of the ISS built off earlier work that modeled how papyrologists read documents by Melissa Terras (Terras 2005). This model identified various levels of reading conducted by papyrologists (identifying features (strokelets), characters, series of characters, morpheme, grammatical level, meaning of word, meaning of groups of words, meaning of document), but the use of knowledge-elicitation techniques by Terras such as "think aloud" protocols also revealed that "interpretation as a meaning-building process" did not invariably begin at the feature level and then successively build to higher levels of reading. Instead, as Tarte explained, the creation of interpretations jumped between levels of reading, and interpretations at any given level might influence interpretations at another. Roued-Cunliffe articulated this point further:

> The conclusion drawn from this experience was mainly that reading ancient documents is not a process of transcribing the document letter-by-letter and line-by- line. Instead it is a cyclic process of identifying visual features and building up evidence for and against continually developing hypotheses about characters, words and phrases. This is then checked against other information in an ongoing process until the editors are happy with the final interpretation (Roued-Cunliffe 2010).

Of particular interest for the development of their ISS then, was to determine "how and why the jumps between reading levels occur, and to what extent vision, expertise and interpretation are intertwined" (Tarte 2010).

Further insight into this process came from the examination of the transcript of three papyrologists attempting to figure out a complicated letter-form. Two major approaches were identified, a "kinaesthetic/paleographical approach" where the scholar would draw characters or trace over them with a finger to try and reconstruct the movements of a scribe, and a "philological/cruciverbalistic" approach where the scholar looks at the question as a puzzle solving task and often relies upon characters they are certain of to make decisions and try out various hypotheses (Tarte 2010). Although Tarte recognized that the two approaches were not mutually exclusive, she also concluded that the ISS would need to be able to support both approaches. Through analyzing this transcript she also identified several forms of scholarly expertise and how they triggered jumps between reading levels, including visual skill (from experience), "scholarly content expectations" (based on prior knowledge), aspect-shifting ("ways of looking" vs. "ways of seeing,") and global-local oscillations. Translating the work processes that lead to scholarly interpretations into digital methods, however, Tarte stated has not been a simple task. "One difficulty in building a case for an interpretation is that it is all about reconstructing a meaning for which there is no accessible ground truth," Tarte reported, "The objective towards which we are tending is to facilitate the digital recording of how such a case is made" (Tarte 2010). The project has thus sought to unravel the process of making decisions and to "mind map" various percepts (through the creation of a schematic of percepts) that lead to the creation of interpretations.

Additional technical details on design choices for the ISS that would support the cruciverbalistic/philological approach have also been given by Roued-Cunliffe (2010). She explained that since DUGA needs to record not just final scholarly decisions but the evidence used to create them, she has explored the idea of "using a set of knowledge bases, such as word lists and frequencies from relevant corpora" to suggest interpretations of words and letters as a scholar reads a document or as evidence to support a particular interpretation. Consequently, DUGA includes a "word-search facility" that is connected to a "knowledge base Web Service" called APPELLO that has been created from the EpiDoc XML files of the Vindolanda tablets. The project also hopes to further develop the knowledge base function of this web-service to support any textual corpus that uses EpiDoc. Rather than creating a large rulebase for classicists such as those created in a DSS system for doctors, Roued-Cunliffe noted that classicists often use similar documents to make choices about the one they are currently analyzing, so the use of a knowledge base from related documents seemed a far better choice.

The work of Roued-Cunliffe differed slightly from the earlier work of Melissa Terras in that she identified "stages" of reading rather than levels and placed certain identification tasks (such as the identification of letters) on a different level. Her work still relied on the basic idea that interpretations consist of "networks of percepts" from low to high level, and that these percepts are used to make scholarly decisions in an iterative fashion. Roued-Cunliffe also wanted to make sure the model used for DUGA reflected the actual working practices of classicists, and did not simply rely on quantitative measures:

> Classical scholars do not traditionally justify their interpretations for example by claiming to be 85% sure of a character or word. Therefore, there would be no point in trying to quantify their perceptions by expressing a percentage of certainty for a given percept. Instead this research is working on a model of evidence for (+) and against (-) each percept. The network of percepts would furthermore enable each percept to act as evidence for or against other percepts (Roued-Cunliffe 2010).

Thus Roued-Cunliffe wanted to make sure that DUGA could capture scholarly expertise and add reasons used by scholars to make decisions as "pieces of evidence" under the heading "Scholarly Judgments." The current DUGA prototype stores ongoing interpretations as XML documents, but does not yet store all the pieces of evidence for and against each percept.

The basic DUGA prototype is divided into a set of views, a "transcript view" and a "box view" that visualizes each character and word with boxes, both of which are populated by two different XSLT translations of an XML document. Since users will view images of documents and need to make annotations on words or characters, Roued-Cunliffe is exploring integrating the image annotation tool AXE that is currently being enhanced by the TILE project (Porter et al. 2009) or an annotation viewer created by the BVREH project (Bowman et al. 2010). The use of one of these annotation tools would allow users to draw "character-, word- or line-boxes anywhere

on the image at any time" and these annotations would then be turned into XML and could have scholarly arguments attached to them. Scholars could also move back and forth between the box view and the annotation view, as they needed to add or review characters or words. The ability to "support a circular interpretation process" Roued-Cunliffe argued was an essential design feature for DUGA. As more scholars use such a system to annotate texts, the evidence used for and against different percepts such as problematic character identifications could be stored and then presented to new scholars as they were annotating the same text. In addition, the eSAD project is also working on character recognition software that will also make recommendations. "The word search and character recognition results should not be seen as conclusive evidence," Roued-Cunliffe concluded, "but as suggestions that may either confirm the scholars' current percept or inspire a new one. It is entirely up to the scholars to decide how they value each piece of evidence" (Roued-Cunliffe 2010).

In order to extend their model, Roued-Cunliffe commented that more knowledge bases would need to become available, such as a knowledge base of Greek words for a scholar working on inscriptions. The general idea would be to allow scholars to choose from a number of relevant knowledge bases depending on the text with which they were working. Their current web service APPELLO makes use of the highly granular encoding of the EpiDoc XML files of the Vindolanda project, particularly the lemmas that were encoded. In the future, they hope to adapt APPELLO so that it can interact with other classical language datasets available online, and Roued-Cunliffe hoped that more projects would make their datasets available in a format such as XML. Other planned work is to turn the prototype into a working application, where the biggest issues will be designing an interface, determining how to store the networks of percepts and the evidence for and against them, and adding annotation software.

Philology

A recent article by Moalla et al. (2006) has defined philology as a research field that studies "ancient languages, their grammars, the history and the phonetics of the words in order to educate and understand ancient texts" and that "is mainly based on the content of texts and concerns handwriting texts as well as printed documents." Crane, Seales and Terras (2009) similarly define philology as the "production of shared primary and secondary sources about linguistic sources" and distinguish classical philology as a discipline that "focuses upon Greek and Latin, as these languages have been produced from antiquity through the present."

As these definitions illustrate, the study of philology concerns all texts whether they are ancient manuscripts or printed editions from the nineteenth century. The needs of philologists are closely tied to the development of digital editions and digital corpora, and various research surveyed throughout this review has explored different facets of philological research. For example the LDAB helps philologists find the oldest preserved copies of individual texts and portals such as KIRKE and Propylaeum have created selected lists of digital philological resources. The project TextGrid is dedicated to creating a specialist text editing environment and philologists are one of their intended user groups (Dimitriadis et al. 2006, Gietz et al. 2006). Various computational tools such as morphological analyzers, lexicons and treebanks have also been developed to assist philologists of Sanskrit (Huet 2004, Hellwig 2010), Latin (Bamman and Crane 2006, Bamman and Crane 2009), and Greek (Bamman, Mambrini and Crane 2009, Dik and Whaling 2009). Other tools have also been created to help philologists create digital critical editions such as DUGA (Roued-Cunliffe 2010), Hypereidoc (Bauer et al. 2008), and OCHRE. In addition, other work has explored cyberinfrastructure for digital philology and digital classics (Crane, Seales and Terras 2009). This section, however, will look at several research projects that hope to support a new type of "digital philology."

One of the greatest obstacles to "digital philology" according to some researchers is that digital corpora such as the TLG and the PHI simply choose a *single* edition as their canonical version of a text and provide no access to the apparatus criticus (Boschetti 2009, Ruhleder 1995):

> Such approach to the ancient text, just about acceptable for literary and linguistic purposes, is unfeasable for philological studies. In fact, the philologist needs to identify manuscript variants and scholars' conjectures, in order to evaluate which is the most probable textual reading, accepting or rejecting the hypotheses of the previous editors. Furthermore, he or she needs to examine the commentaries, articles

and monographs concerning specific parts of the text. Thus, the extension in breadth of the aforementioned collections needs to be integrated by the extension in depth, according to the paradigms of a new generation of digital libraries (Boschetti 2009).

In the Digital Aeschylus project described by Boschetti (2009), the author reports that they are seeking to remedy these problems by creating a digital library that includes images of multiple manuscripts of Aeschylus, manually created transcriptions of the most relevant manuscripts and printed editions, OCR of recent editions, an extensive bibliography of secondary sources, and information extraction tools to be used on the digitized documents. They seek to create a comprehensive digital library for Aeschylus that will support philologists in the development of critical editions.

Tools for Electronic Philology: BAMBI and Aristarchus

One of the earliest projects that explored the computational needs of philologists was the BAMBI ("Better Access to Manuscripts and Browsing of Images") project that developed a "hypermedia workstation" to assist scholars in the reading of manuscripts, writing annotations, and navigating between words in a transcription and images in digitized manuscripts (Bozzi and Calabretto 1997). The project was aimed at two types of users, general users of libraries that wished to examine manuscripts and "professional students of texts" or philologists, who they defined as "critical editors of classical or medieval works that are hand-written on material supports of various types (paper, papyrus, stone)" (Bozzi and Calabretto 1997). The authors thus developed a "philological workstation" that included four major features: 1) the ability to look up digital images in an archive 2) the transcription, annotation and indexing of images 3) the viewing of transcribed versions of texts and creating an "Index Locorum" 4) the automatic matching of words found in transcriptions, the "Index Locorum" and annotations with the relevant portion of the source document image that contains the word. Interestingly, this last feature while desired by many other digital edition and manuscript projects, is still an area of unresolved and active research (Cayless 2008, Cayless 2009, Porter et al. 2009).

In an overview of their philological workstation, Bozzi and Calabretto provided a list of the functions that it supported. To begin with, the workstation allowed users to search manuscript collections and to create transcriptions of digital images of manuscripts and export them as RTF or SGML. One important feature was the indexing of transcriptions that could be used by philologists to generate an "Index Verborum" and an "Index Locorum" for each script in the manuscript (e.g. Greek and Latin). This "Index Verborum" contained *all* the words appearing in the transcription and the words that were *corrected* by the user (using the text variant function), while the "Index Locorum" displayed "the positions in which each word occurs in the manuscript." In addition, annotations could be created on manuscript transcriptions, and all annotations contained two distinct fields, one for free comments and the critical apparatus, and one for variants, synonyms, and the correction of syntax. The BAMBI workstation also supported automatic column and line recognition, and even more importantly the automatic creation of a word-image concordance (if a transcription for a manuscript was available) that matches each word of the text with the appropriate portion of the image. The concordance was built automatically and this module provided a simultaneous view of the transcription and the image so the user could check its accuracy and it also allowed the user to query the manuscript collection by selecting a word in either the transcription or on the image. The BAMBI prototype made use of HyTime (an extension of SGML) to model works on ancient manuscripts, in particular because it allowed "specification of links between text and part of image (part of an object)."

While the fuller technical details of this workstation are somewhat outdated as of this writing, the unanswered issues identified by the BAMBI project are still largely relevant for digital philology today. Bozzi and Calabretto noted that the following requirements needed to be met: better standards based tools for the description of manuscripts, more sophisticated image processing routines (although they called for the enhancement of microfilm images rather than the images of manuscripts themselves), "a comprehensive solution for the management of text variants," "tools based on image processing facilities and linguistic (statistical) facilities for the electronic restoration of missing text elements," new models for collaborative work (though work today has moved beyond client-server models based on the web), and a survey of the technical and legal issues involved in creating "widespread, multi-source services offering digital versions of library materials and

the tools for their use" (Bozzi and Calabretto 1997). As has been seen in this review, the challenges of manuscript description, advanced image processing, the management of text variants, the creation of sophisticated digital tools, collaborative workspaces, and comprehensive open-source digital libraries all remain topics of concern.

Other research in digital philology has been conducted by the Aristarchus project [429] and a recent article by Franco Montanari (Montanari 2004) has provided an overview of the electronic tools for classical philology available at the website. Montanari suggested that two types of digital tools had been created for philology and indeed for the humanities in general: 1) general electronic tools that were transformed to fit more specific needs 2) new tools that were created to meet unique demands. According to Montanari, an increasingly familiarity with digital tools would be required of all philologists: "The "new" classical scholar and teacher is supposed to be at home with this kind of tools," Montanari asserted, "Textual, bibliographical, and lexicographical databanks represent three of the most relevant electronic tools available thanks to the progress of digital technology."

The Aristarchus project, named after Aristarchus of Samothrace, includes a number of tools for philologists studying the Greek and Roman world and has been created by the University of Genoa. The first tool, the "Lessico dei Grammatici Graeci Antichi" (LGGA) [430] or "Lexicon of Ancient Greek Grammarians" provides a lexicon of ancient Greek scholars and philologists and provides an online database that can be used to study the "history of ancient philology, grammar and scholarship." In addition, this website provides access to a second "lexicon" the "Catalogus Philologorum Classicorum" (CPhCL) [431] "an encyclopaedic lexicon that collects the biographies and the bibliographies of modern classical scholars." The largest database is "Poorly Attested Words in Greek (PAWAG)" [432] and it gathers together Ancient Greek words that have only been rarely attested and is described by Montanari as a "half way house between a dictionary in the strict sense and an encyclopedic lexicon." Two specialist websites have also been created by Aristarchus: MEDIACLASSIC [433] (a "web site for didactics of the ancient Greek and Latin languages") and "Scholia Minora in Homerum" [434] (a site that provides an "up-to-date listing, descriptions, editions and digital images of the so-called Scholia Minora to the Iliad and Odyssey on papyrus"). Images of papyri can be viewed after registration. Finally, the Aristarchus website also hosts the Centro Italiano dell'Annee Philologique (CIAPh), [435] the Italian editorial office of the international Année Philologique.

Infrastructure for Digital Philology: the Teuchos project

While both the BAMBI and Aristarchus projects explored the use of digital tools, both projects defined philology in a fairly traditional manner in terms of the type of work that would be performed. A more expansive definition of philology was offered by Crane et al. (2009b): "philology is thus not just about text; it is about the world that produced our surviving textual sources and about the tangible impact that these texts have had upon the worlds that read them." To pursue a new level of ePhilology, the authors argued that a new digital infrastructure needed to be developed that brought together all relevant primary and secondary sources that are currently scattered in various specialized digital libraries and to provide background knowledge personalized to the needs of individual scholars. In addition, new digital editions and commentaries need to abandon the limited assumptions of print publications (e.g. simply scanning a printed book rather than creating a true digital edition), Crane et al. (2009b) reasoned:

> We now face the challenge of rebuilding our infrastructure in a digital form. Much of the intellectual capital that we accumulated in the twentieth century is inaccessible, either because its print format does not lend itself to conversion into a machine-actionable form or because commercial entities own the rights and the content is not available under the open-licensing regimes necessary for eScience in general and ePhilology in particular (Crane et al. 2009b).

[429] http://www.aristarchus.unige.it/index_inglese.php
[430] http://www.aristarchus.unige.it/lgga/index.php
[431] http://www.aristarchus.unige.it/cphcl/index.php
[432] http://www.aristarchus.unige.it/pawag/index.php
[433] http://www.loescher.it/mediaclassica/
[434] http://www.aristarchus.unige.it/scholia/
[435] http://www.aristarchus.unige.it/ciaph/index.php

Thus the lack of machine-actionable contents and restrictive copyright regimes frustrate a move to ePhilology. In addition, a cyberinfrastructure for ePhilology the authors argued required at least three types of access: 1) "access to digital representations of the human record" such as page images of manuscripts and printed books, 2) "access to labeled information about the human record" such as named entity annotations, and 3) "access to automatically generated knowledge" or the processes of various algorithms.

Creating such a new digital infrastructure for philological research is one of the larger goals of Teuchos, [436] a project of the University of Hamburg in partnership with the Aristotle Archive at the free university of Berlin. Teuchos is building a research infrastructure for classical philology, with an initial focus on the textual transmission of Aristotle. Work will focus on the digitizing, encoding and description of manuscripts, developing a XML encoding for manuscript watermarks, and creating a web-based environment for philological work that includes a Fedora repository, the management of heterogeneous data, and support for a multi-lingual editing environment. Two recent articles (Deckers et al. 2009) and (Vertan 2009) have explored different aspects of the Teuchos project.

Deckers et al. (2009) offered a detailed explanation of the data encoding and representation of "manuscripts as textual witnesses and watermarks" with a focus on the former and an extensive overview of the Teuchos platform.

In its final form Teuchos is to provide a web based research environment suited for manuscript and textual studies, offering tools for capturing, exchange and collaborative editing of primary philogical (sic.) data. The data shall be made accessible to the scholarly community as primary or raw data in order to be reusable as source material for various individual or collaborative research projects. This objective entails an open access policy using creative commons licenses regarding the content generated and published by means of the platform (esp. digital images of manuscripts may have to be handled restrictively dependant upon the holding institutions' policies) (Deckers et al. 2009).

The Teuchos project is consequently developing an open source platform that can be used for collaborative editing of manuscripts and creation of philological data, and the data that is created will be made available under a CC license, although they noted that providing access to images of manuscripts will depend on the respective policies of their owning institutions. [437] One particularly distinctive feature of the Teuchos platform is that it will support the integration of heterogeneous data and the participation of different user groups.

The creators of Teuchos (Deckers et al. 2009) outlined a number of potential use cases that informed their design choices, including: 1) the provision of extensive data that facilitates the use of digitized manuscripts such as the markup of both the *structural information* and *intellectual content* of manuscripts (this would include transcriptions that indicate variant readings); 2) access to digital manuscript images that are accompanied by at least *partial* transcriptions so that material not only becomes more rapidly available and citable but can also be the "basis for further editorial work"; 3) a collaborative environment for researchers; 4) a constantly evolving collection of manuscript descriptions that provides scholarly information on "codicology, manuscript history and textual transmission"; 5) a flexible data model that can accommodate the integration of "manuscript descriptions" of varying semantic depth and length; and 6) linking to important existing online resources such as library catalogues, specialist bibliographies, and digital texts. Deckers et al. (2009) also reported that they particularly wanted to create a tool that provides scholars in the fields of Greek codicology and palaeography with the ability to publish digital research materials.

The Teuchos platform is built off of a Fedora repository and three types of users can interact with this repository through a web application, [438] systems administrators, registered users that can also contribute resources, and public users that can only view publicly released materials. The Teuchos Fedora repository includes several types of complicated digital objects all of which have been designed to try and cover all potential categories of text transmission. Manuscript watermark tracings are stored as digital images and information about them is stored in a custom XML format created by the project. A "textual transmission" group has two subgroups that are each then subdivided further, the first group provides information related to individual *manuscripts* while the

[436] http://beta.teuchos.uni-hamburg.de/
[437] Copyright, creative commons licensing and the use of digitized manuscript images has recently been explored by (Cayless 2010a).
[438] A beta-version of this application is now available http://beta.teuchos.uni-hamburg.de/TeuchosWebUI/teuchos-web-ui

second includes information related to individual *works*. Within the manuscript group, individual data objects include digital page images (of either complete or partial manuscripts) that are aggregated for each manuscript, codicological descriptions that reference page images when available, and varying levels of transcription data. In terms of *works*, this subgroup encompasses a wide range of materials referring "to a source text with its entire set of manuscripts rather than to one particular witness" and includes full critical editions, translations, and commentaries (Deckers et al. 2009). The three other major categories of digital object that are also created are biographical dictionaries, bibliographical data and published research papers. Due to the heterogeneous nature of this data, only the manuscript descriptions and transcriptions were able to be encoded according to TEI P5 XML.

As the creators of Teuchos hope to provide scholars with advanced searching and editing functionality, they have developed a data model for both the physical and intellectual content of manuscripts in their platform. While not all of the descriptive material in Teuchos includes digital images of manuscripts, all the digital images that are included have accompanying descriptive and authority metadata. All manuscripts with digital images also have a corresponding reference document that makes use of the TEI <facsimile> element and a list of <surface> elements with unique identifiers in the form of xml:id attributes and unambiguous labels for pages using the "n" attribute. The <surface> elements are listed in the physical order of the manuscript and missing pages are represented with empty <surface> elements.

In order to facilitate user access to individual page images, Teuchos also provides at least a minimal transcription for each manuscript (e.g. it may simply contain page break information and no textual transcription) that contains structural information that "can be used to offer alternate representations and improved navigation for browsing, and to give a clearer indication of the part of the text to which an image viewed pertains" (Deckers et al. 2009). This data is then encoded within TEI <text> elements. While <pb> elements with "corresp" attributes that point to unique page identifiers are used to reference digital images of individual manuscript pages, the <fw> element is used to separately encode foliation or pagination information. This separate encoding is important, as Deckers et al. reported because it "permits recording whether numbers provided by the transcriber are actually present on the page or not" and also supports "recording more than one such reference system," a particularly important issue, since many manuscripts can have multiple foliation systems.

A comprehensive set of markup structures has also been created to represent the intellectual content of manuscripts. Deckers et al. (2009) observed that two complementary issues are involved in relating structural information to a transcription, first, the need to relate the text of a manuscript transcription to the structure found in a *particular edition* of a work, and second, the need to encode "any structure evident" in the actual manuscript witness that is being transcribed. These issues become even more complicated, Deckers et al. argued when combining the transcriptions of multiple manuscript witnesses of a work:

> To be able to retain per-witness structural information in a joined document, we therefore propose to encode all structural information using empty elements, i.e. <milestone>s. When such a joined document is edited further to become a new edition of a work in its own right, the editor(s) may (and in most cases probably will) of course decide to create a hierarchical structure taking into account the structure of the various witnesses, but this should be a later step. To avoid confusion, we should state that we do not intend to provide dynamically generated editions. While the semi-automatic joining of transcriptions is a first step towards creating a digital critical edition, the further steps require substantial scholarly intervention.

The approach chosen by Teuchos illustrates the difficulty inherent in trying to create dynamically generated editions, particularly for those works that have potentially dozens of manuscript witnesses. In addition, it has often been the case that structural information from an older existing edition is used as the *organizational* structure for a new edition (e.g. using Stephanus edition page numbers for an OCT edition of Plato). Deckers et al. thus proposed a hierarchical system that used <milestone>s with a "special value of "external" for the unit attribute" that made it clear an external reference system was being indicated and a specially created value of "canonical" for the type attribute. Different edition and numbering schemes were referred to using an "ed" attribute, the hierarchical level of the reference used a subtype attribute, and the actual reference then used a "n attribute." Defining this level of encoding granularity is important for it means that *multiple* canonical reference

systems from different editions or even multiple numbering schemes from one edition can be encoded for one manuscript text. This same system is also used to encode the *content* structure of individual manuscript witnesses, but a value of "internal" is used instead of "external" for the unit attribute with type values of "present" for numbers actually found within the text and "implied" for numbers that are no longer found in the witness. Finally, Deckers et al. also suggested that the "ed" attribute could be used to indicate manuscript siglum as well as edition names.

While future work for the Teuchos project involves the creation of detailed codicological manuscript descriptions and transcriptions of individual manuscript texts, Deckers et al. explained that they first focused on structural encoding, since they considered "this an important step in providing fuller access to digitised manuscripts for textual scholars, and a necessary prerequisite for cumulative and shared scholarly work on the primary text sources in a distributed digital environment." As with the work of Interedition and the Virtual Manuscript Room, the Teuchos project wants to create a distributed environment that will let many scholars contribute their expertise and share their editing work with others. A recent presentation by Vertan (2009) has offered some further technical details on this infrastructure that is being built, and stated that Teuchos is working with CLARIN as one of their collaborative research projects: "MLT-Cphil-Multilingual Language Technology for Classical Philology Research." Vertan described Teuchos as a "Knowledge Web- Based eResearch Environment" where knowledge work is supported through knowledge organization, semi-automated information extraction, the management of multilingual data, and "intelligent retrieval of heterogeneous materials"; where the use of the Web will allow for comprehensive data access (different levels of users), interoperability (TEI P5) and persistency (URIs and persistent identifiers), user modeling, and the creation of a shared workspace; and where the "eResearch environment" provides access to different material types, sophisticated data modeling, encoding, and visualization and extensive linking between different digital projects (Vertan 2009).

As part of their collaborative environment, Teuchos also plans to create a shared workspace that includes a forum and to support various commenting and versioning features for different materials. Perhaps the greatest challenge, however, listed by Vertan was the need to manage both multi-lingual and heterogeneous data that included different data types such as semi-structured data found in XML and TEI files, high-resolution TIFF images, graphics (for watermarks), and research materials stored as PDF or Word documents. The semi-structured documents also had varying levels of semantic depth and there were different types of multilinguality such as within one document (e.g. Greek, Latin and German in one manuscript), across documents (there are five official languages for the project: German, French, English, Italian, Spanish) and within terminologies. The Teuchos project wants to allow navigation across different data collections and they are currently implementing various Semantic web solutions. This includes "semantic descriptions of stored objects" using RDF triples, developing ontologies for each type of data collection, mapping "multilingual lexical entries" onto this ontology and then supporting ontological searching.

Managing all of this data involves the creation of complicated digital objects in Fedora that have 7 data streams: bibliographic details (stored in DC), semantic descriptions of objects and relationships with other objects (RDF), codicological descriptions (XML), linguistic information (XML), layout details (XML), transcriptions (text file), image data (TIFF files). The current Teuchos implementation not only makes use of Fedora but also uses AJAX for the client-server application and image viewer. Vertan concluded that the Teuchos platform could illustrate the potential of semantic web technologies for real humanities problems as well as demonstrate the importance of developing solutions for multilingual content. In addition, Vertan asserted that "multilingual problems are increased due to the lack of training data and CL tools for old languages, especially ancient Greek." Similar criticisms were offered by Henriette Roued who noted the lack of Ancient Greek knowledge bases for use with the DUGA prototype.

Prosopography

The discipline of prosopography is "the study of individuals" and in terms of ancient history "is a method which uses onomastic evidence" or the study of personal names "to establish (i) regional origins of individuals and (ii)

family connections."[439] Many different sources can be used for prosopography including narrative texts, administrative records, letters, and inscriptions among many others. The study of prosopography has thus been closely linked to epigraphy in particular. This section will look at several recent articles and research projects that have investigated the use of digital techniques in creating prosopographical databases.

Issues in the Creation of Prosopographical Databases

Although prosopography is a well-established discipline, there are fewer digital resources in prosopography than in many of the other fields of digital classics, an issue that has been discussed extensively in a recent study by Ralph W. Mathisen (Mathisen 2007) who provided an overview of existing prosopographical databases (PDBs) and the challenges involved in creating them.[440] In his own early work in the 1970s, Mathisen decided to create a database based of the first volume of the *Prosopography of the Later Roman Empire* (PLRE), but had to temporarily abandon this work due to the limitations of FORTRAN and mainframe computers. By the 1980s, Mathisen believed that the development of PDBs was becoming increasingly possible and in one of his own grant proposals at the time listed a number of major advantages of databases, including convenience, speed, accuracy, diversity and multiplicity of access, ease of revision and reporting, expandability, portability, and perhaps most importantly, potential compatibility with other biographical and prosopographical databases. His current work involves the development of a database he has named the "Biographical Database of Late Antiquity"(BDLA).[441] Despite the potential of PDBs, Mathisen reported how his earlier research (Mathisen 1988) had identified 20 prosopographical database projects, but by 2007, only one had been completed, one had been absorbed by a later project, two were still in progress, and the other sixteen were no longer findable.

A variety of issues have caused this situation according to Mathisen including questions regarding accessibility and hardware and software problems, but the greatest challenges have been methodological considerations of the discipline. To begin with, Mathisen noted that some prosopographers (and indeed other humanists) argued that databases imposed "too much structure on the information from primary sources." Any useful historical database, Mathisen suggested, must structure data from primary sources in two ways, first it must identify all "categories of recurrent information" (e.g. sex, religion) and second, it must identify "appropriate, recurrent values for these fields." Mathisen also pointed out that "historians *always* structure their data, whether they are creating a PDB or not." While creating a database may be an *interpretative* act of scholarship as earlier argued by Dunn (2009) in terms of archaeology, *using* a historical database also involves interpretation as explained by Mathisen:

> PDB structure and coding are not prescriptive; it only provides a starting point for research. The computer can only do so much. Human intervention is always needed, not only in the course of the creation of a PDB, but especially in the use of a PDB. This includes not only verifying the validity and appropriateness of the data returned, but also judiciously analysing that data. Even when I read an entry in the hard-copy of *PLRE*, I still check the primary source myself. Users of PDBs should do the same (Mathisen 2007).

As this statement illustrates, Mathisen considered scholarly consultation of the original primary sources to also be extremely important. To be truly useful, however, Mathisen also proposed that PDBs should include at least some access to the primary sources they used whenever possible. "Indeed, the most effective modern PDBs bring the original source documents along with them," Mathisen argued, "either by a pointer to a separate source database or by including the source text within the record, thus ensuring that no source information is ever lost in the creation of a PDB"(Mathisen 2007).

Another major methodological issue in the development of PDBs according to Mathisen is that they are about "individual people" and as such these people must have unique identities within a database. Yet the

---

[439] "prosopography" *Oxford Dictionary of the Classical World*. Ed. John Roberts. Oxford University Press, 2007. *Oxford Reference Online*. Oxford University Press. Tufts University. 4 May 2010.
<http://www.oxfordreference.com.ezproxy.library.tufts.edu/views/ENTRY.html?subview=Main&entry=t180.e1853>

[440] In this overview, Mathisen also notes the somewhat limited research coverage of PDBs in the last ten years, with much of the significant scholarship published in the 1980s, such as (Bulst 1989, Mathisen 1988) and the 1990s (Goudriaan et al 1995).

[441] As of this writing (September 2010), there does not appear to be any website for the BDLA, which according to (Mathisen 2007), "plans to incorporate all the persons attested as living in the Mediterranean and western Asian worlds between AD 250 and 750" and contains over 27,000 individuals and includes over 70 searchable fields.

identification of individual people within primary sources is no easy task, and even if two sources cite the person with the same name it can be very difficult to determine if it is the same person. Additionally, individuals can go by different names. "Sorting out who's who," Mathisen noted, "either by using a computer algorithm or by human eye-balling, continues to be one of the major problems, if not the major problem, facing the creators of PDBs" (Mathisen 2007). As has been seen throughout this review, the challenges of historical named entity disambiguation have also been highlighted in terms of historical place names in archaeology (Jeffrey et al. 2009a, Jeffrey et al. 2009b) and classical geography (Elliott and Gillies 2009b) and both personal and place name disambiguation complicated data integration between papyrological and epigraphical databases in the LaQuAT project (Jackson et al. 2009).

While hierarchical structures were first explored for PDBs, Mathisen proposed that it was generally agreed that the relational model was the best structural model for such databases. Several important rules for relational PDBs that Mathisen listed included: the need to store data in tabular format, the creation of unique identifiers for each primary data record (within PDBs this is typically a person's name combined with a number, e.g. Alexander-6), and the importance of the ability to retrieve data in different logical combinations based on field values. While many PDBs, Mathisen observed, were often "structured based on a single table" that attempted to include all the important information about an individual, such a simple structure limited the types of questions that could be asked of such a database.

The final major methodological issue Mathisen considered was that of standardization. While the early period of PDB creation saw a number of efforts at developing a "standardized format for entering and storing prosopographical material," Mathisen doubted that any real standardization would ever occur. Indeed, he argued that since the "data reduction" methods of any prosopographical database were often designed based on the primary source material at hand and how it would be used, attempting to design an all-purpose method would be inefficient. While Mathisen proposed that the use of a relational database structure in itself should make it relatively easy to transfer data between databases, the LaQuAT project found this to be far from the case (Jackson et al. 2009).

Despite these various methodological issues, a number of prosopographical database projects have been created, as shall be seen in the next sections. Mathisen posited that there were two general types of PDBs: [442] 1) a restricted or limited database that typically incorporates individuals from only one "discrete primary or secondary source" and 2) "inclusive" or open-ended databases that usually include all of the people who lived at a particular time or place and contain material from many heterogeneous sources. As will be seen from this review, all of the databases considered in the following sections, with the exception of Prosopographia Imperii Romani, are open-ended databases. Such databases are far more difficult to design, according to Mathisen, since "designers must anticipate both what kinds of information users might want to access and what kinds of information will be provided by the sources from which the database will be constructed." In addition, such databases are typically never completed as new resources become unearthed or additional sources are mined for prosopographical data. "The greatest future promise of PDBs lies in the construction of more sophisticated and comprehensive databases," Mathisen concluded, "Including a broad range of persons, constructed from a multiplicity of sources and permitting searching on a multiplicity of fields" (Mathisen 2007).

Network Analysis & Digital Prosopography

In 2009, a new digital research project entitled the Berkeley Prosopography Service (BPS), [443]received funding from the Office of Digital Humanities (ODH) of the NEH to create "an open source digital toolkit that extracts prosopographic data from TEI encoded text and generates interactive visual representations of social networks." [444] This project is led by Niek Veldhuis, along with Laurie Pearce and Patrick Schmitz, and they are

[442] Mathisen also lists a third special case of limited databases with the form of open-ended databases, but that "are constructed from existing hard-copy prosopographical catalogue" or card-files, and the limit is imposed not by source-material but by editorial decisions on whom to include. In addition, Mathisen also described a number of "biographical catalogues" like the "De Imperatoribus Romanis" (DIR) (http://www.roman-emperors.org/).
[443] http://code.google.com/p/berkeley-prosopography-services/
[444] http://www.neh.gov/ODH/Default.aspx?tabid=111&id=159

utilizing both NLP and social network analysis (SNA) techniques to extract personal names and family relationships of people mentioned in texts and to then assemble a social network of people based on described activities.[445] The initial tool will be applied to a corpus[446] of approximately 700 cuneiform tablets from the CDLI that record sales and lease transactions among a small group of Mesopotamians from Uruk (southern Iraq) between 331-346 BCE. After the Uruk text corpus[447] has been converted into TEI-XML, prosopographic data will be automatically extracted from the TEI files, SNA techniques will be used to create various networks, and users will then be able to visualize the results in various ways:

A probabilistic engine will collate all the person-references in the corpus, along with some basic world knowledge, like the typical length of adult activity, and will then associate the names to individual persons, and finally will relate the people to one another by the kind of activities they engaged in. The resulting graph model can be used to produce a variety of reports and visualization tools, including simple name lists and family trees, as well as interactive models. By integrating graph visualization tools, the project will provide interactive tools that let researchers explore the network of associations and activities. They can focus on an individual, on a given type of activity (e.g., real-estate sales), or explore other aspects of the model. This should enable the researchers to answer many complex questions more easily, and with a visual response (Schmitz 2009).

The BPS will provide researchers with individual workspaces and will also be the first independent tool to be incorporated into the CDLI. During this initial grant period beta-testing will also be conducted with other corpora to test both scalability and generalizability of this tool for use in other prosopographical projects. The BPS is also participating in Project Bamboo as one of their demonstrators.

Other work using network analysis[448] as a means of exploring classical prosopography has been discussed by Graham and Ruffini (2007). They noted that rapid developments in computer technology, particularly graph theory, provided both network analysts and consequently prosopographers with new tools for answering complex questions involving the degrees of separation between network members or how densely or loosely connected networks might be:

Such questions hold a natural interest for prosopographers, who can then begin to look for certain characteristics—class, office, occupation, gender—and identify patterns of connectivity that they might have otherwise missed when confronted with a mass of data too large for normal synthetic approaches. And yet, network analysis has been slow to take root among ancient historians. Network analytical research on the Greco-Roman world has focused on questions of religious history and topography. Nonetheless, the epigraphic and papyrological evidence beg a network analytical approach to the prosopographical data available from these sources (Graham and Ruffini 2007, pg. 325-326).

In order to demonstrate the value of network analysis for prosopography, the authors described their own dissertation work. One major requirement they listed that would be needed to demonstrate the potential of network analysis for ancient prosopography were "focused data-sets" unlike many of the massive multi-volume prosopographies such as the PLRE.

As an example of one such data set, Graham and Ruffini described a set of data regarding individuals connected with the brick industry of imperial Rome. This data was largely obtained from bricks that were stamped with estate and workshop names and all together this data set included the names of at least 1300 individuals from largely the 2nd century A.D. As individuals involved in the brick industry came from varying levels of society, the name data from the bricks has been used in various types of historical research. Several major published catalogues of stamped bricks have been created and Graham created an Access database for one of them (CIL XV.1) that could be used for both archaeological and prosopographical analysis. Numerous programs can then be used to build and analyze networks from this data Graham and Ruffini suggested:

In general, one simply lists the name in question and all the other names with which it co-occurs. The programme then stitches the network together from these data. Many statistics of use to prosopographers can then be determined, but sometimes simply visualizing the network itself can provide a 'eureka' moment. Some networks will have a number of 'hubs' and everyone else is connected like a

[445] http://inews.berkeley.edu/articles/Spring2009/BPS
[446] The demonstrator corpus "Hellenistic Babylonia: Texts, Image and Names (HBTIN)" can be viewed at http://oracc.museum.upenn.edu/hbtin/index.html.
[447] http://cdl.museum.upenn.edu/hbtin/
[448] One of the presentations at the Digital Classicist/ICS Work in Progress Seminar during the summer of 2010 also examined the use of network analysis in prosopography, see Timothy Hill, "After Prosopography? Data Modelling, models of history, and new directions for a scholarly genre." http://www.digitalclassicist.org/wip/wip2010-03th.html

'spoke'; other networks will look more like a chain with interlocking circles of individuals. This is profoundly important (Graham and Ruffini 2007, pg. 328).

Graham consequently used network analysis to explore "small world" networks within Rome and the effect of purges and proscriptions on this network.

Another potential use of network analysis for prosopographical research listed by Graham and Ruffini was for "exploring the interactions between various cliques and clusters within a social network" on the level of individual villages, such as those described in documentary archives of papyri that survive for a number of villages. They noted that the large number of papyrological databases such as APIS and DDBDP provide a wealth of material that can be mined for prosopographical analysis or as they call it "a prosopographical growth industry with enormous potential." The dissertation work of Ruffini used network analysis with documentary papyri from the Aphrodito archive to explore the prominence of individuals other than the heavily studied Dioskoros and his family. Ruffini suggested that network analysis provides a number of "centrality measures" such as "closeness centrality" and "betweenness centrality" that can be used to "identify the most central figures in the archive, measures whose quantitative nature hopefully removes the biases introduced by our own scholarly curiosity and prejudice." Using these two measures identified three other prominent individuals, results that surprised him as none of them are mentioned in modern scholarship on Aphrodito. A final potential use of network analysis for prosopography illustrated by Graham and Ruffini was the analysis of occupational groups and the social connectivity between them.

While Graham and Ruffini acknowledged that most of their analysis is still fairly speculative, they also convincingly argued that the unique nature of their results derived from network analysis of ancient evidence suggests that there are many interesting avenues of future work.

Relational Databases and Modeling Prosopography

Perhaps the most extensive prosopographical database online is the Prosopography of the Byzantine World (PBW).[449] This website, formerly known as the Prosopography of the Byzantine Empire (PBE), provides access to a database that attempts to include details on every individual mentioned in both Byzantine textual and seal sources[450] between 641 and 1261 A.D. The database of the PBW is both large and complex, and as described by the website is composed of thousands of "factoids":

> Its core is made up of nearly 60,000 factoids (small pieces of information classified under different categories), each of which is linked to an owner and (generally) at least one other secondary person by a hypertext link. More than a third of the factoids are of the narrative type, and these are organised into narrative units by further links. There are 2,774 such units. The units are in turn linked to dates and reigns, and some of them to larger events and problems. There are, in addition, around 7,500 seals, with links to matrices which number 5,000. Each seal is linked to a museum or private collection and at least one edition, and each set of matrices to an owner, certain or hypothetical, in the core of the database.[451]

As of April 2010, there were approximately 10,000 named individuals included in the PBW. A variety of searching and browsing options are available for this database. The entire prosopography can be browsed alphabetically, and clicking on an individual name brings up a record for that person that can include varying levels of detail depending on the number and specificity of attestations in the sources. For example, a record for "Kissiane 101" includes a Greek representation of the name, place of residence, and a kinship link to her husband, whereas the record for her husband "Nikephoros 148" also includes a textual description, four kinship relations (all of which are hyperlinked to the records for these individuals), and a list of possessions. Every "factoid" in each individual record also includes the source where this attestation was found. For historically significant individuals, such as emperors, even more extensive sets of factoids are available. For example, the record for "Michael 7" includes 307 narrative factoids, 7 education factoids, and three alternative names among extensive other detail. The PBW offers an extensive level of detail by often including the full text of the various "factoids" from primary sources in each individual record. One feature that is unfortunately not available but

[449] http://www.pbw.kcl.ac.uk/content/index.html
[450] A full list of the primary sources used and their abbreviations is provided (http://www.pbw.kcl.ac.uk/content/index.html). A list of the editions used for the text of the seals can be found at http://www.pbw.kcl.ac.uk/content/reference/sealedit.html
[451] http://www.pbw.kcl.ac.uk/content/reference/full.html

would likely be very useful is the ability to link to individual person records within the PBW with permanent URL's.

While the entire PBW can be browsed alphabetically by individual names a user can also choose to browse the individuals found within individual sources (rather than all) such as the *Alexiad* by Anna Comnena or the *Epitome* by Joannes Zonaras. The user can also choose to browse lists of individuals classified by factoids their records contain (but only one factoid may be chosen at a time) including narrative, authorship, description, dignity/office, education, kinship, language skill, occupation, possession, or religion. In addition to these browsing options, the PDB database search allows the user to keyword search within all factoids, within individual factoids, or within a combination of factoid categories using Boolean operators.

An article by Bradley and Short (2005) has offered some insights into the creation of highly structured databases such as the PBW from sources used in the study of prosopography.[452] As illustrated above, the data in the PBW is drawn from a large number of primary sources, and while Bradley and Short acknowledge that many traditional humanities computing projects might have sought to first create digital editions of these primary sources, they believed that the prosopographical nature of their project required a different solution:

> This is because a digital prosopographical project does not aim to produce a textual edition. If it is to be true to its name, it must create instead, a new secondary source. Like a classic prosopography such as the *Prosopography of the Later Roman Empire*…a digital prosopography must act as a kind of visible record of the analysis of the sources produced by the scholars as they try to sort out who's who from a close analysis of the extant source materials (Bradley and Short 2005).

In traditional printed prosopographies this activity typically results in a biographical article that summarizes what can be concluded about the life of an individual from different sources and interpretative arguments from a scholar as to why they have drawn their conclusions. A distinguishing feature of the PBW, then, as a "new-style" digital prosopography is that its final publication is as a "highly structured database" not as a series of articles.

As Bradley and Short explain and as seen above, all evidence data within the PBW has been recorded as a series of factoids, or assertions made by a member of the project that a "source 'S' at location 'L' states something ('F') about Person 'P'" (Bradley and Short 2005). According to Bradley and Short, a factoid is not a definitive statement of fact about a person and a collection of factoids should not be considered as a scholarly overview of a person. Instead, factoids simply record assertions "made by a source at a particular spot about a person." Since factoids may contradict each other (e.g. different assertions about an individual's ethnicity), all factoids about a person are included in the database. The database also includes a place where prosopographers can record their own assertions about why they have interpreted a text in a certain way. This methodology makes it easier to display the uncertainty inherent in determining "facts" about an individual from complicated primary sources and also illustrates that factoids are also "acts of interpretation by the researcher that gathers them." "The ironic flavour of the name 'Factoid' is not accidental," Bradley and Short submitted, "It reflects the historian's worry when a tiny extract is taken out of the context of a larger text and the historical period in which it was written and presented as a 'fact'"(Bradley and Short 2005). Nonetheless, one difficulty with the factoid approach was how to establish what types of factoids should be collected and historical events proved to be the most challenging kind of data to transform into factoids.[453]

Since factoids link different kinds of structured information together and there were thousands of factoids (60,000 or so according to the website), a relational model was chosen to help users make sense of all of this data. The relational model also offers many new facets for access as most printed prosopographies only offer two to three indices to articles they contain. Bradley and Short contrast their process of creating a database with the "text-oriented modelling" of projects such as the Old Bailey Online.[454] The Old Bailey Online provides access to a searchable online edition of the historical printed proceedings of the Old Bailey, and like most

---

[452] This article also offers some details on the creation of two related database projects, the "Prosopography of Anglo-Saxon England (PASE)", (http://www.pase.ac.uk/pase/apps/index.jsp) and the "Clergy of the Church of England Database" (CCEd) (http://www.theclergydatabase.org.uk/index.html)

[453] The computational modeling of historical events can be very complicated and was also described by (Robertson 2009) in his discussion of HEML.

[454] http://www.oldbaileyonline.org/

prosopographical projects is based on *narrative* texts that include references to people, places and things. While person names are marked up in the XML text of the Old Bailey Online, Bradley and Short remarked that there was no effort "to structure the *names* into *persons* themselves." This is in direct contrast to their relational approach with the PASE, PBEW, and CCEd:

> Our three projects, on the other hand, are explicitly prosopographical by nature, and the identification of persons is *the* central task of the researchers, as it must be in any prosopography. They must have a way to separate the people with the same recorded name into separate categories, and to group together references to a single person regardless of the spelling of his/her name…. It is exactly because prosopographical projects are involved in the creation of a model of their material that is perhaps not explicitly provided in the texts they work with that a purely textual approach is in the end not sufficient in and of itself. Instead, it is exactly this kind of structuring which makes our projects particularly suitable for the relational database model (Bradley and Short 2005).

In addition, the databases for all three of these projects contain not only "structured data in the form of factoids" but structures that are spread over several tables and represent other important objects in the database including "persons, geographic locations, and possessions."

Bradley and Short also addressed a point raised earlier by Mathisen regarding the limitations of historical databases and the interpretation and categorization of data. As Mathisen maintained, they argued that all work with prosopographical sources, whether writing an article or creating a database involved a fair amount of scholarly interpretation and categorization. Rather than attempting to create an "appropriate" model of their sources, Bradley and Short argued they were trying to create a model of how prosopographers *work* with those sources:

> For, of course, our database is not designed to model *the texts* upon which prosopography is based with all their subtle and ambiguous meanings. The database, instead, models the task of the prosopographer in interpreting them i.e. it is not a model of an historical text, but a model of *prosopography* itself (Bradley and Short 2005).

The importance of modeling how scholars within a discipline conduct their work and how they work with their sources are important components in the design not just of historical databases but also in larger digital infrastructures that will need to support multi-disciplinary work.

Mathisen has described the approach of the PBW as a combination of a "multi-file relational model" with a "decentralized biography model" (Mathisen 2007) or where instead of having an individual record with dedicated fields created for each individual, each person is instead assigned a unique ID key that is then associated with the information bites or "factoids" as described above in various other databases. "Biographies" are thus created for individuals by assembling all the relevant factoids for an individual. Mathisen offered a few caveats in terms of the methodology chosen for the PBW, namely, that the complexity of the data structure would make it hard for anyone without expert computer skills to implement such a solution and that the "multiplicity of sub-databases and lack of core biographies" would make it difficult to export this material or integrate it with another PDB without specialized programming (Mathisen 2007). He also feared that the lack of "base-level" person entries might mean that important information for individuals could be omitted when different factoids were combined and could also make it difficult to determine when occurrences of the same name represent the same or different individuals. Despite this assertion, Bradley and Short proposed that by not providing their users with an "easy-to-read" final article about each individual and instead presenting a collection of "apparently disconnected and sometimes contradictory factoids" they are in fact bringing the user closer to doing actual prosopographical work. They argued that the series of factoids could be read as a "proto-narrative" and also serve to remind users of the interpretative and fuzzy nature of the data that they are getting from the database. Bradley and Short also asserted that the PBW seeks to provide focused access to the primary sources themselves and that users of the PBW should also consult these sources to form some of their own interpretations. Thus the importance of access to primary sources is illustrated again in the study of prosopography.

Other Prosopographical Databases

A major prosopographical resource for ancient Greece is Website Attica, [455] an online database that has been designed to complement and extend a series of published volumes entitled *Persons of Ancient Athens* (PAA). Additions and corrections that are made to the published volumes are also included in the online database. Over 10,000 Athenian names are included in the database and a large variety of searching features are available. Individual names must be entered in capital letters in Greek transliteration. As the website explains, possible searches "range from selecting every person in a particular deme or of a specified profession to more sophisticated searches" such as finding "all Athenians who lived between specified years and/or are related to a certain person and/or are attested in a class of document." The record for each individual includes an identifier and identified name and may also include the following: status, place (a field which contains the "demotic or ethnic of a person"), phyle, link (kin relationship), kin name, activity, date, and a comment field where all additional information about a person that did not fall into one of the above categories can be found. A separate bibliographic reference search of the database is also available.

One online resource for Roman prosopography is the Prosopographia Imperii Romani (PIR), [456]a website that is maintained by the Berlin-Brandenberg Academy and provides an online index to the person entries found in the printed volumes of the *Prosopographia Imperii Romani*. The first edition of this series was published in three parts between 1897 and 1898, and a second edition was published in seven parts with multiple fascicules beginning in 1933 and concluding in 2006. The individuals covered in the PIR are drawn mainly from the upper levels of society (emperors, senators, knights and their related family members) of the Roman Empire between 31 B.C. and the end of the reign of Diocletian (284-305). The source material used in creating both the printed volumes and this database is wide-ranging and includes literature (Ovid, [457]Virgil, Plutarch, Horace, Pausanias), administrative and historical records as well as inscriptions, papyri and coins. Access to the PIR entries is provided through a searchable "keyword list" that has been created for the website, and each entry contains a unique identifier, a person's name, and a reference to the printed PIR volumes or other standard reference works.

A variety of ongoing research into the "onomastics and prosopography of the 'later' periods of Egyptian history on the basis of the Greek, Latin, Egyptian and other texts" is being conducted by researchers using the various texts contained within the Trismegistos portal. [458] The basic methodology involves the collection of anthroponyms and toponyms mentioned in the texts, and when there is no electronic corpus available, these names are entered manually. Work with Greek papyri, however, has been greatly enhanced due to the existence of the XML encoded corpus of the DDBDP, which has been made freely available to them. Since the DDBDP is in Unicode and has already capitalized all proper names, the extraction of names from it was greatly simplified. The names extracted from the DDBDP were added to the list of personal names already available from the Prosopographia Ptolemaica, and currently the full corpus includes "25723 Greek nominative name variants" that have been grouped into 16571 names. Links from this merged corpus to the DDBDP will be made using a database of 207,070 "declined forms of these name variants." Ultimately, all of the recognized name forms will be stored in a relational database of name references that will then be able to serve as a prosopography. All name references will be linked to the appropriate texts in the Trismegistos texts database.

The Prosopographia Ptolemaica (PP) [459] is a long-standing research project from the department of Ancient History at the University of Leuven. While this project first started as a "list of all inhabitants of Egypt between 300 and 30 B.C., from Greek, Egyptian and Latin sources" it has recently been extended to include the Roman and Byzantine periods. This resource has been integrated into the larger Trismegistos portal and includes close links to the HGV and the DDBDP but also maintains a separate database interface. This database can be searched by Latin name transcription, ethnic group, residence, PP number or date and each individual person

[455] http://www.chass.utoronto.ca/attica/
[456] http://www.bbaw.de/bbaw/Forschung/Forschungsprojekte/pir/de/Startseite
[457] One interesting project created an onomasticon exclusively for the *Metamorphoses* of Ovid http://staff.cch.kcl.ac.uk/~wmccarty/analyticalonomasticon/
[458] For a list of the projects see, http://www.arts-humanities.net/event/digital_classicistics_work_progress_seminar_onomastics_name_extraction_graeco_egyptian_papyri_
[459] http://ldab.arts.kuleuven.be/prosptol/index.html

record can include a PP number (if available), a Latin transcription of the name, sex, place of residence, ethnic group, assumed dates, and a reference to the text in which they were mentioned (e.g. papyri, inscriptions) along with a link to bibliographic information on this text in the Trismegistos database.

Another website that provides access to prosopographical data from Egypt is the website "DIME Online: Prosopgraphie zu Soknopaiu Nesos." [460] DIME contains references to written records (Demotic and Greek) of people who lived in the Soknopaiu Nesos area of Al Fayyūm from the 7th century B.C. to the fifth century A.D. The entire database can be searched and each identified individual has a descriptive record that includes basic personal and kinship information, possessions, and any relevant bibliography. While searching of the database does not require registration, if a user registers they can also add information to the database.

A related *onomastic* if not entirely prosopographical project for Ancient Greece is the Lexicon of Greek Personal Names (LGPN). [461] This project was first established in 1972 as a research project of the British Academy under the direction of Peter Marshall Fraser, and in 1996 it became a part of Oxford University and is a member of the group of Oxford Classics Research Projects. The purpose of the LGPN is to:

> collect and publish with documentation all known ancient Greek personal names (including non-Greek names recorded in Greek, and Greek names in Latin), drawn from all available sources (literature, inscriptions, graffiti, papyri, coins, vases and other artefacts), within the period from the earliest Greek written records down to, approximately, the sixth century A.D. [462]

This lexicon does not include mythological names, Mycenaean names, or later Byzantine names. Currently five volumes have been published with several more that are forthcoming. Individual volumes include all the Greek names from a particular geographic area (e.g. LGPN I: Aegean Islands, Cyprus, Cyrenaica). Each individual volume can also be downloaded as a series of four PDF files, with an introduction, a bibliography of sources used and their abbreviations, and a forward and reverse index of the Greek names in that volume. All of the LGPN data (250,000 published records) is stored in a relational database and each record typically includes a normalized primary name form, sex of the individual named, place and date of attestation (dates can vary widely), and the bibliographical reference or references as to where this name was found. [463] This website also includes a useful introduction to Greek names including their history, formation and meanings and an image archive that includes tombstones, vases, inscriptions and other sources that have been used for names. A searchable database called the "LGPN Online" can be used to search the over 35,000 names published in LGPN I-IV and the revised LGPN II. Work is also currently underway to develop a TEI XML schema for the LGPN and to convert the entire database into TEI-XML for long-term preservation and interoperability.

A recent presentation by Matthews and Rahtz (2008) has provided extensive details on the future plans of the LGPN regarding TEI-XML, how the resource has already been used in various types of classical research, [464] and how it may be used in future research. As Matthews and Rahtz described, the LGPN has lived through various generations of humanities computing since it first originated in the 1970s. The most important part of this long history was the development of a database in the 1980s that was "structured to reflect and provide access to all the research components of an LGPN record, which in the books are subsumed under name-headings" (Matthews and Rahtz 2008). While this database has been important in enforcing some format consistency and was used to generate the printed volumes, Matthews and Rahtz also argued that its research potential has yet to be fully exploited. [465] The last decade of LGPN development has involved reaching the following goals: the serialization of the relational database into XML, the support of online searching using an XML database, and a new data model that will emphasize collaboration.

The future plans of the LGPN are to convert their electronic lexicon into a system entirely based on TEI-XML. This work is being undertaken not only to create an IT infrastructure that will support the preservation and

---

[460] http://www.dime-online.de/index.php?PHPSESSID=hc6fl0v16ls14vesuptn7uqnc3
[461] http://www.lgpn.ox.ac.uk/
[462] http://www.lgpn.ox.ac.uk/project/index.html
[463] http://www.lgpn.ox.ac.uk/online/documents/TEIXML
[464] The LGPN has hosted two international conferences regarding its use and the results have been published (http://www.lgpn.ox.ac.uk/publications/index.html) in two separate books. The topics covered included linguistics, religious history, and demographic studies among many others.
[465] For a fuller description of this database and the conversion of the printed slips, see http://www.lgpn.ox.ac.uk/online/computerization/.

maintenance of the LGPN data but also to enable this data to play a larger role in an e-research environment and to allow the LGPN to play a "central role in determining standards for encoding names in documents" through TEI/XML and thus achieve greater interoperability with digital resources worldwide (Matthews and Rahtz 2008). This "XML phase" of the LGPN work has led to the definition of a customized TEI-XML schema that will be used to preserve an archival form of the lexicon data in a digital repository. This work also both coincided with and thus influenced the TEI's recent revision of their module relating to names and dates. [466] The new module models "persons, places and organizations as first class objects" so the LGPN schema is thus a fully "conformant pure subset of the TEI" (Matthews and Rahtz 2008). The LGPN has also usefully defined five potential levels of data interchange: character interchange, character encoding, standardized structural markup, standardized semantic markup, and information linking.

Currently the LGPN has created an experimental database [467] that contains an XML version of the LGPN in a single XML database and uses XQuery to join name, place and person data together to support new forms of sophisticated searching. They are currently delivering search results as HTML, TEI-XML, and KML for use in Google Maps and Google Earth, Atom feeds for use in RSS readers, and JSON (for use in Simile Timeline [468] and Exhibit [469]). Even more importantly, they are providing consistent "cool URLs" so that this data can both be linked to and be widely reused in other applications. Only a limited number of the 3000 attested place names in the LGPN currently have KML downloads since this format requires latitude and longitude for locations, but the LGPN is currently working with Pleiades (as part of the Concordia initiative) to find these places in the Barrington Atlas and utilize the geo-location information found within this atlas. All of these formats Matthews and Rahtz explain are created by a simple series of XSL transformations on the TEI XML file. Through the use of consistent standards, therefore the LGPN was able to successfully demonstrate the potential of linking their data with many other digital classics projects.

As this overview of prosopographical and onomastic resources illustrates, there are a number of prosopographical resources online, although far fewer than for other classical disciplines. None of the projects reviewed it appears other than the LGPN have plans to provide XML versions of their data, or indeed to provide any access to their data at all other than through the individual websites. The information found within these databases, however, particularly the lists of personal names and their variants could be extremely useful as training data in the development of named entity disambiguation algorithms for historical texts. Similarly, as many of the resources used in the creation of these databases (e.g. published collections of documentary texts papyri, inscriptions, etc.) are in the public domain and may have been published online (e.g. in Google Books or the Internet Archive), individual name records could likely be linked to a variety of online sources of their attestations.

The Use and Users of Resources in Digital Classics and the Digital Humanities

While this review originally intended to include a survey of studies that examined how scholars made use of specific digital classics projects and how well they met their needs, no such overview studies were located. [470] A number of digital classics resources included extensive bibliographies [471] that listed research that had made use of the *analog* sources (e.g. printed collections of inscriptions or papyri, published editions of classical texts), but none seemed to include studies that specifically examined either if or how the digital resources were being used

[466] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html
[467] The experimental database can be accessed here http://clas-lgpn2.classics.ox.ac.uk/
[468] The SIMILE Timeline is an open source "widget" that can be used create interactive timelines (http://www.simile-widgets.org/timeline/) and LGPN made particular use of TimeMap (http://code.google.com/p/timemap/) a javascript library that was created to "help use Google Maps with a SIMILE timeline."
[469] Exhibit (http://www.simile-widgets.org/exhibit/) is an open source publishing framework created by the SIMILE project that can be used to easily "create web pages with advanced text search and filtering functionalities, with interactive maps, timelines, and other visualizations."
[470] While some research has been conducted into the use of the Perseus Digital Library, none has been conducted in the last ten years or in terms of the current website (Perseus 4.0), for earlier research, see for example (Marchionini and Crane 1994).
[471] See for example, the bibliography of publications related to Projet Volterra, http://www.ucl,.ac.uk/history2/volterra/bibliog.htm.

for scholarship. [472] There are many studies that investigate the information *seeking* habits of humanities scholars including how they find electronic resources or search on the Web, but many of these studies have focused on "traditional" electronic resources such as databases subscribed to by libraries or the use of general search engines such as Google. [473] One notable exception that focused specifically on humanist use of primary resources, including digital facsimiles, is Audenaert and Furuta (2010), which will be examined in further detail below.

As the discipline of classics falls within the larger umbrella of the humanities disciplines and this review has primarily focused on open access digital resources in classics, a number of studies that investigated the use of freely available digital resources in the humanities have been chosen for further examination to see what general insights might be determined. For this reason, this review has examined various studies that have explored the citation of electronic resources in classics (Dalbello et al. 2006), the behaviors of digital humanists with e-texts (Toms and O'Brien 2008), humanist use of primary source materials including digital facsimiles (Audenaert and Furuta 2010), the scholarly creators of digital humanities resources (Warwick et al. 2008b), and the "traditional" scholarly use of digital humanities resources (Harley et al. 2006b, Brown and Greengrass 2010, Meyer et al. 2009, Warwick et al. 2008a).

Citation of Digital Classics Resources

One method of exploring how digital resources are being used within a discipline is to determine how many citations to different digital resources can be found within "conventional" publications. Pursuing the traditional task of bibliometrics to examine how and when digital resources are cited in scholarly publications is a growing area of research, and a general methodological approach is described as part of the JISC funded project "Toolkit for the Impact of Digitised Scholarly Resources (TIDSR)." [474] In their efforts to determine how often five individual digital projects had been cited, the project team searched for citations to these resources using Google Scholar, [475] Scopus, [476] and ISI Web of Knowledge. [477] One major issue they reported was that bibliometrics "with regard to non-traditional scholarly outputs is that citation habits in many fields favour citing the original paper version of a document, even if the only version consulted was electronic." [478] In fact, in their own study they found that of those scholars who published papers as a result of their work with digital materials in the five projects, over 1/3 of them only cited the physical item that was represented in the collection and made no reference to the digital project at all, almost half cited the original publication but also included the URL, and less than one in five cited only the digital version. Thus they cautioned against simply relying on bibliometrics to analyze the actual scholarly impact of a digital project. As project director Eric Meyer explained:

> This means that relying on finding citations to one's digitised resource based on looking for URL's within journal citations is almost certainly going to yield an artificially low number because of the uses that don't cite it at all, and because of inconsistencies in how the URLs are cited. Nevertheless, doing regular searches for citations to a collection's material is an important way to establish the impact it is having on the scholarly community. [479]

Thus while one way to measure the impact of a resource in digital classics is to perform a citation analysis looking for citations to project URLs or references to digital projects in article text using various tools such as Google Scholar, the actual amount of use of different digital projects may be quite higher than can be easily

---

[472] For some collections such as the APIS, relevant bibliography of how an individual papyrus has been used or published is integrated into records for the papyri.

[473] A synthesis of major findings from over twelve recent studies in this area (including faculty, researchers, graduate students and undergraduates from various disciplines) has been created by Connaway and Dickey (2010) while Palmer et al. (2009) have offered an analysis of studies that have focused more exclusively on the scholarly practices and information use of faculty online during the last twenty years. See also the literature review found in (Toms and O'Brien 2008). For two sample recent examinations of how humanists search for information on the Web and work with library electronic resources see (Amin et al. 2008) and (Buchanan et al. 2005)

[474] http://microsites.oii.ox.ac.uk/tidsr/kb/53/bibliometrics-enhancing-ability-track-projects-scholarly-impacts

[475] http://scholar.google.com/

[476] http://www.scopus.com/

[477] http://www.isiknowledge.com/

[478] http://microsites.oii.ox.ac.uk/tidsr/kb/53/bibliometrics-enhancing-ability-track-projects-scholarly-impacts

[479] "What is Bibliometrics and Scientometrics?" Eric T. Meyer. http://microsites.oii.ox.ac.uk/tidsr/kb/48/what-bibliometrics-and-scientometrics

determined. Meyer's point also illustrates the importance of maintaining stable URL's to both encourage the citation of resources and their subsequent discovery in bibliometric analyses.

As there are dozens of digital classics resources mentioned in this paper, a full bibliometric analysis of even one of the projects would be far beyond the scope of this review. Nonetheless, sample searches within Google Scholar for two projects, the Prosopographia Ptolemaica and the APIS, illustrated that these resources are indeed cited within the larger scholarly literature, even though these citations are not always easy to find. For example, the Prosopographia Ptolemaica has been used to explore spatial relationships and estimate population size settlements (Mueller and Lee 2004), as one source of data for an onomastic study of Hebrew and Jewish-Aramaic names (Honigman 2004) and as a source of data for a population study of Hellenistic Egypt (Clarysse and Thompson 2006). Digital images and translations of individual papyri within the APIS have also been cited in different publications, including a discussion of a comic fragment (Barrenechea 2006) and in the study of Greek hepatoscopy (Collins 2008). Interestingly, while all three of the citations to the Prosopographica Ptolemaica either listed the database by name in the article or footnotes or included URLs to the collection, both of the references to the APIS didn't include URLs to the individuals papyri they utilized. [480]

A specific citation study for electronic resources in classics was conducted by Dalbello et al. (2006) and examined both the number and types of citations to electronic resources made by classicists in three important journals (*Classical Journal*, *Mnemosyne* and *Classical Antiquity*). [481] As the authors considered classics to be a field known for digital innovation they expected to find many references to digital scholarly resources, but instead found that references were typically made to educational sites such as in articles that discussed learning about the discipline, in reports of practice that discussed the recent history of the discipline, and in articles that analyzed the potential use of technology for research. "More rarely are the associations to electronic resources included for knowledge building in a traditional scholarly fashion," Dalbello et al. observed, "such as – integrated in the literature review, supporting the main argument, etc." Within classical journals the types of websites cited included community of practice sites, university sites, digital library sites, encyclopedias, online dictionaries, electronic libraries and electronic journals. Thus Dalbello et al. remarked that while digital resources were discussed as teaching tools or in state-of-the-art reviews, they were not being used to create new knowledge or being used as research tools. Despite the presence and use of digital resources, Dalbello et al. concluded that most classicists still perceived of publication as paper-based:

> Our findings indicate that the structuring of literature in these fields is largely still perceived as paper-based….Documentary cultures resulting from digitization of resources supporting traditional research and digital preservation as well as multiple document formats for scholarly journals (electronic, paper) present a new research environment for the humanities disciplines that is not as yet fully integrated in the canonical knowledge base. These citation practices point to the still invisible nature of the electronic document that is now ubiquitous in supporting the actual research practice (Dalbello et al. 2006)

The lasting influence of the print paradigm has also been explored in terms of "digital incunabula", or how the conventions and limitations of print publication have also shaped the development digital projects (Crane et al. 2006).

The Research Habits of Digital Humanists

Another approach to determining the actual use of digital classics resources is to survey the users who work with them. Since no studies that specifically examined the scholarly use of digital classics materials could be found, a recent comprehensive study that investigated in detail the research habits of digital humanists with e-texts was used to gain insight into larger research and use patterns of digital humanities scholars and how these might reflect the behaviors and needs of those scholars who use digital classics resources.

A study by Toms and O'Brien (2008) focused in particular on self-identified e-humanists and how they utilized ICT in order to inform the design of an "e-humanist's workbench." Their research results were based on small sample of digital humanists who responded to a web-based survey, and Toms and O'Brien noted that they

[480] But the record for the papyrus utilized in (Barrenechea 2006) does list his publication in its bibliography (http://wwwapp.cc.columbia.edu/ldpd/apis/item?mode=item&key=columbia.apis.p1550)
[481] This same research also explored electronic citations made in English literature journals

planned to expand their research with interviews and actual observations of scholars at work. As part of their research they examined dozens of articles and over 40 years of studies on scholarly information behavior and the "traditional" research habits of humanists (e.g. with printed library materials or databases). One fundamental conclusion they reached from this overview was that: "The accumulated research depicts the humanist as a solitary scholar who values primary materials and secondary materials – namely books – and engages in browsing behaviour more than searching" (Toms and O'Brien 2008). They also observed that the "information seeking strategy of choice" was linking rather than searching or a combination of browsing different materials and chaining, or using one relevant article to find other articles (e.g. through the footnotes or by who has cited the article in hand). Their overview of the accumulated literature also illustrated that when humanists began research they cared more about depth than relevance and this helped to facilitate the development of ideas, the ability to make connections, and the creation of an initial knowledge base to which later knowledge could be related.

One problem that Toms and O'Brien found with the various theoretical studies of information behavior [482] that they considered was that they typically excluded how information is actually *used* by humanists after it is found and thus are of limited use in the development of actual systems. "Despite the profound impact of technology on this scholarly community," Toms and O'Brien remarked, "little is known about how computers have affected humanists' work flow, unless it is to say that scholars adopt technologies when they augment established research practices"(Toms and O'Brien 2008). Nonetheless, earlier work conducted by one of the authors of this paper (Toms and Flora 2005) had identified a concrete set of needs for e-humanists that including five key components: 1) access to primary sources; 2) presentation of text; 3) text analysis and relevant tools; 4) access to secondary sources; 5) tools for communication and collaboration. In addition, Toms and O'Brien also noted that there is little joint publication in the humanities. Having thus reviewed both the work of information scientists and humanists, they elaborated on three common themes: 1) humanities scholarship utilizes a diverse set of primary and secondary sources and while text is the primary resource, a variety of digital media are also used; 2) digital humanists use a variety of tools and techniques when working with encoded texts; 3) humanists were typically solitary researchers (e.g. they saw little evidence of joint publication) but did communicate with other scholars.

Seeking to either confirm or expand these findings, Toms and O'Brien thus conducted a survey of self identified e-humanists and in order to expand their knowledge regarding the current use of electronic texts by e-humanists, the research environment of e-humanists, and the types of research performed by all humanists. Participants for the survey were recruited through listservs such as HUMANIST [483] and all results were obtained from a questionnaire that asked about general information, teaching and research interests, and the use of ICT. The survey results of the 169 respondents were then analyzed using several different software tools. The ratio of male to female survey respondents was about 3:2, almost 2/3 of respondents had PhDs, and the majority were from Canada followed by the United States, Europe and Australia. Of those surveyed, approximately 25-30 percent had never performed any type of text analysis. In terms of areas of specialization, over 40% reported that they worked in literature, but "classics, history and religion" was reported by 12% of participants. While the "dominant language" that most scholars reported working in was English, a dozen other languages were also identified and interestingly more scholars reported working primarily with Latin (17%) than Italian (7%) or Spanish (9%). Another survey question examined the historical period in which scholars mainly worked and while more than half worked primarily with materials from the 20th century, approximately 13% worked in ancient and "post-classical history."

As teaching and research are often intertwined in the humanities, Toms and O'Brien also asked respondents how they used electronic texts and ICT in the classroom. About 60% of their respondents taught courses in the humanities, but their use of technology in the classroom overwhelmingly involved the general use of course websites, online instructional systems, and textbook courseware. While in some cases students were required to

[482] As an overview of all of these studies Toms and O'Brien provided a table that summarized that various stages of the information seeking process identified by different researchers.
[483] http://www.digitalhumanities.org/humanist/

create e-texts (39%), encode texts using markup such as HTML or XML (72%) and use text analysis tools (33%), Toms and O'Brien also reported that "a significant number of respondents (42 per cent) have not required students to use any of these." Thus although many e-humanists used text analysis and digital technology in their own research, there was an apparent disconnect for many in terms of their use of technology in the classroom.

Another set of questions asked by Toms and O'Brien explored the research themes of e-humanists. The most prevalent research theme listed was the semantic or thematic examination of the text/s of one or more authors (37%), this was followed by the creation or use of specific electronic editions (20%) and the creation of specialized catalogues or bibliographies using databases that already exist (13%). Only 13% reported their main research theme as conducting "computational text analyses" or "developing techniques for analysis."

Questions regarding the use of electronic texts and text analysis tools made up a large body of the survey, and 86% of respondents reported using e-texts. At the same time only 34% had used "publicly available" text analysis tools and 61% had never scanned or encoded texts. Interestingly in terms of markup, over half of the respondents stated that they preferred no markup in their e-texts and almost 25% had no knowledge of TEI. Electronic texts were typically selected according to levels of access, special features and cost, and in general scholars wanted texts to be legally accessible, available in a stable form and from reliable publishers or institutions. Similarly, over 75% wanted e-texts to be peer-reviewed (with 67% preferring texts from established editions), while 79% wanted e-texts to be accompanied by documentation (79%). Surprisingly, less than half of those surveyed (48%) required page images to be available. In addition, only 62% of respondents had used text analysis tools, and those who had not used them reported various reasons including expense, low priority, usability issues and technical incompatibility.

When asked about their "wish list" for text analysis tools, two of the most significant desires listed were for institutions "to maintain tools for the study and publication of e-texts" (41%) and for fellow researchers to share tools they had created (46%). While 66% of participants had either created or contributed to the creation of text analysis tools, most respondents were unaware of currently available text analysis tools. Of those that were aware of currently existing tools, respondents typically considered the majority of them to not be very useful. Desired text analysis techniques were also quite varied but the two most frequently desired capabilities were the to ability to compare two or more documents (69%) and to view a text concordance (61%). In sum, a large number of e-humanists desired to have some type of institutional infrastructure for their work, but also displayed a lack of knowledge about what types of resources were already available.

Another series of questions gauged participants' access to primary and secondary sources. Over 90% of respondents rated search engines as highly useful for finding e-texts and analysis tools and over 78% wanted to be able to view lists of available e-texts. Survey respondents also wanted a reasonably high level of structure for their e-texts, since 71% wanted to be able to restrict their search terms by chapter, with 53% wanting to restrict it by a character in a play or novel and 48% wanted to search on the level of the individual paragraph. These results are interesting since many participants also reported that they preferred no markup in their texts, and searching at these levels of granularity requires at least basic *structural* markup (e.g. chapters, pages, paragraphs) and in the case of novel characters, *semantic* markup (e.g. TEI).

The final series of questions involved scholarly communication and collaboration and the large majority of answers seemed to confirm the picture of humanists, even self identified e-humanists, as solitary researchers. [484] As Toms and O'Brien reported, almost half of respondents worked alone. In addition, a majority had not conducted research with colleagues (55%) or graduate students (64%). An even larger number of researchers (87%) reported that they did not tend to discuss their work before it was formally submitted, with less than 40% sharing ideas at early stages of research and over half acknowledging that they had not consulted colleagues at

[484] Similar conclusions were reached by Palmer et al. (2009) in their overview of online scholarly information behavior across disciplines: "Thus, humanities scholars and other researchers deeply engaged in interpreting source material rely heavily on browsing, collecting, rereading and notetaking. They tend to compile a wide variety of sources and work with them by assembling, organizing, reading, analyzing and writing. In interacting with colleagues, they typically consult rather than collaborate, with the notion of the lone scholar persisting in certain fields"(pg. 37)

all. While their research had confirmed the picture of the humanist as a solitary scholar, the authors also proposed that this was perhaps due more to the nature of work in the humanities rather than personal qualities:

> This does not, however, mean that humanists are not collegial; it may be more fitting to say that humanists communicate with each other rather than collaborate, since collaboration implies working together – building – and the humanists' work is all about deconstructing ideas and dissecting texts (Toms and O'Brien 2008).

To try and facilitate greater collaboration in the future, Toms and O'Brien suggested that a e-humanist workbench should provide a variety of communication and collaboration tools.

In analyzing their findings, Tom and O'Brien concluded that this encapsulated view of digital humanists at work illustrated that "clearly, humanities research is intricate and diverse." They were surprised both by the relatively low level of technology use within the classroom and by the fact that for many respondents the use of technology simply involved delivering reading materials from a course website. Another notable finding according to Toms and O'Brien was that search engines were used as often as library catalogues to locate both primary and secondary sources, a finding that marked a significant change from many of the earlier studies they had found. Library tools were typically used for well-defined topics and browsing remained a preferred method for finding information. As a result of these findings, they decided that an e-humanities workbench should include a web search capability as well as links to catalogues, finding aids and archives. A scholar should also be able to personalize the workbench with their own list of relevant web sites and digital libraries.

The authors also offered a number of conclusions regarding workbench support for e-texts and text analysis tools. To begin with they stated that any workbench should support the downloading, storing and organizing of e-texts as well as encoding them in different markup languages. "With the multiple forms of mark-up and the multiple expectations about the multiple mark-ups," Toms and O'Brien observed, "it is clear that "multiple views" of a text are needed for e-humanists" (Toms and O'Brien 2008). In addition, as both availability and access to texts are critical to the work of humanists, Toms and O'Brien argued that part of the problem is difficulty not just in identifying what texts have been digitized but also in gaining *access* to these texts. Since UNIX tools had been reported as among the most useful, Toms and O'Brien also reasoned that the workbench would need to support both the awareness and use of already existing text analysis tools through both technical and peer support and examples of what the tools can do. "Access to text-analysis tools is imperative" Toms and O'Brien acknowledged, "But more importantly, the development of new tools is badly needed by this community. While TEI and TEI-lite standardized the mark-up work for humanists, similar standards need to be developed to serve as the basis for text analysis tools, so that text with various forms of mark-up are interchangeable with different types of tools"(Toms and O'Brien 2008). The authors thus make the important point that standards are needed not just for text markup but for the development of compatible text analysis tools as well.

The results of Toms and O'Brien study illustrate some important themes to be considered when designing a digital infrastructure for classics, namely the need to provide a common framework/research environment where scholars can both find and use already existing text analysis tools and e-texts, to support granular scanning and browsing tools so that scholars can interact with a text, and to include sophisticated text analysis and annotation tools for use with texts in a variety of markup languages. Additionally, such an environment should also include communication and awareness technologies so scholars can communicate regarding projects and resources and also share tools and research methodologies.

Humanist Use of Source Materials: Digital Library Design Implications

A recent study by Neal Audenaert and Richard Furuta has offered in-depth insights into how humanists use original source materials (both analog and digital facsimiles) as well as made a number of recommendations for the design of humanities digital libraries. While Audenaert and Furuta conceded that a large number of digital humanities resource collections already exist, they argued that most of these collections sole purpose was to *disseminate* materials. "Digital libraries, however, hold the potential to move beyond merely disseminating resources," Audenaert and Furuta explained, "toward creating environments that support the analysis required

to understand them. To achieve this, we must first develop a better understanding of how humanities scholars (and others) use source documents in their research" (Audenaert and Furuta 2010). In order to create a digital library that could provide an environment for research as well as disseminate resources, Audenaert and Furuta undertook a user study as part of a larger research program to design a "creativity support environment (CSE) to aid in-depth analysis and study of paper-based documents."

The authors argued that there is an urgent need for this type of research because the "interdisciplinary digital humanities field" has been largely dominated by the humanities community, where scholars have developed resources that simply meet their own *individual* research needs or that have embodied "theoretical" definitions of what digital scholarship should be. They concluded that these issues make such resources and methodologies of limited use for informing *large-scale* systems design for the broader humanities community:

> This work tends to focus on describing the objects of study from within the framework of a specific theory, rather than the more traditional human-centered systems approach of analyzing the goals of specific user communities and the tasks they use to achieve those goals. The resulting tools may do an excellent job of supporting the humanities scholars' needs for "thick description" but often result in work practices that are intimidating to many scholars (for example, the expectation that scholars will manually encode documents using XML) or that emphasize topics such as authorship attribution that are far from the mainstream of humanities research (Audenaert and Furuta 2010).

In addition, Audenaert and Furuta also reiterated an earlier point made by Toms and O'Brien regarding user studies from library and information science, that such studies are typically focused on *information retrieval* rather than *information use*. Thus their work sought to characterize how scholars actually *used* materials that they found and examined in detail how access to materials supported their research questions as well as considering what types of insights scholars gained from original materials and what if any use they might make of a CSE. In a series of semi-structured interviews with eight scholars, Audenaert and Furuta asked scholars both why and how they worked with original sources through a series of open-ended questions that focused on three research themes: 1) Why do scholars put in the time to work with original materials? 2) What information are they looking for? 3) How and when do they use computers, or do they use them at all? The scholars they interviewed worked in a variety of humanities fields but also included two scientists who focused on the use of historical documents.

To begin with, they learned that scholars put in the effort of using original source materials for a variety of reasons including: many original sources and transcriptions are now easily available, to form "holistic impressions" and gain a sense of a text as a physical object, to examine objects/sources in "nuanced detail," to alleviate their concerns with the accuracy or authenticity of a transcription or edition (many scholars did not want to trust the work of others and didn't always trust their own work without notes), and for the "aesthetics" of working with original documents. Audenaert and Furuta stressed that even though in many cases transcriptions were considered *adequate*, there were still many times when scholars insisted that access to *originals* (digital or analog) would be essential:

> While editors will try to identify and describe relevant details in their published editions, the level of detail required, the specificity required by different lines of research, and the need for visual inspection makes it impractical to describe all of this information in secondary sources. Consequently, many lines of inquiry require access to source material (either directly or through digital surrogates) even when high-quality editions are readily available (Audenaert and Furuta 2010)

This point echoes the earlier discussion of Bodard and Garcés (2009), who argued that open source critical editions should provide access to all their source materials, so that scholars can form their own conclusions and ask new questions.

Findings for their second major research topic that examined what scholars were looking for in original documents illustrated four major themes. Scholars were interested in textual transmission, surveying all the evidence and documents on a topic, identifying all the agents who contributed to both the creation and transmission of a text (e.g. author, audience, editors, illustrators, publishers, scribe), and documenting the full social, political and economic context of text. Interestingly, the study of text transmission, a critical task of much classical scholarship, was the most common goal of all the scholars surveyed. Another critical point made by Furuta and Audenaert was that since many scholars wanted to "survey all documentary evidence" related to a

text or particular topic (a task they noted that had been made more "tractable" by modern editions), cultural heritage digital library designers should consider a "systematic survey of a collection of source documents" when creating digital libraries.

The third major research question explored how scholars used computers and whether they would be willing to use digital study and research tools in their work. The CSE that Audenaert and Furuta contemplated creating would include support for both "information seeking and externalizing formative ideas." While information seeking has received a great deal of attention, Audenaert and Furuta noted that "externalizing knowledge" had received less attention. [485] As the humanities research process often involves intimate experience with both the secondary literature and disciplinary methods of a field, they stated that such knowledge is both "implicit and voluminous" and individuals are often unwilling to formally express such knowledge (e.g. through an ontology). At the same time their study also reflected that participants kept both detailed and systematic notes and that they usually kept them electronically. They defined the scholars' process of research as "incremental formalism" where they were focused on a specific final product such as a monograph or scholarly article. Furuta and Audenaert thus asserted as have many other studies cited in this report (Bowman et al. 2010, Porter et al. 2009) that scholars' would benefit from both note-taking and annotation support and also from a comprehensive digital environment that supported all the steps of the research process from formal notes to a final publishable manuscript.

One major conclusion Audenaert and Furuta drew from this research was that while digital facsimiles were not "adequate for all research tasks" they nonetheless played a critical role in "mediating" access. In general, they noted that scholars' were most concerned with the editorial contributions they made to a digital project, and the major form of computational support that was desired was for tools that could help them prepare and disseminate their work to the larger scholarly community. While Audenaert and Furuta acknowledged that scholars were not particularly reflective or "critically oriented" to their own work practices, they still believed that computational support for *all* levels of the research process should be provided. "To the contrary, we would suggest that the clear (and relatively easy to achieve) benefits of applying technology to support the dissemination of scholarship, coupled with the comfortable familiarity of existing disciplinary methods," Audenaert and Furuta articulated, "has led the digital humanities community to overlook opportunities to critically assess how new technology might be developed to support the formative stages of scholarship" (Audenaert and Furuta 2010).

Audenaert and Furuta argued that scholars' work with original source materials forms part of a "complex ecosystem of inquiry" that involves both understanding a text and its full context of creation, transmission and use. They defined this process through the SCAD model, which consists of five components: primary objects that are studied, the multiple sources of a document (e.g. previous drafts or copies), context (historical, literary, political, etc.), actors, and finally, "derived forms" or the related sources for which an original text may in turn serve as the source for (e.g. text reuse and repurposing of content). The main goal of designing the SCAD model the authors explained was to serve "analytical goals" and to be used as a *tool* that could guide designers developing tools for scholars rather than a formal conceptual model for "representing information in humanities digital libraries." Somewhat in contrast to the work of Benardou et al. (2010a), Audenaert and Furuta were uncertain as to whether scholars would be willing to use tools that formally represented structures and relationships between information.

A related point offered by Audenaert and Furuta was that cultural heritage digital libraries and repositories need to re-conceptualize their potential roles and move beyond serving primarily as final repositories for scholarship, to also serve as resources that can support *research* that is in process. Another important insight they offered was that since many humanities digitization projects can take years that digital libraries need to be designed as "evolving resources" that support the "entire life cycle of a research project" from the digitization of materials to ongoing research using those resources to the publication and preservation of long-term scholarly works.

[485] Some preliminary work on externalizing the methods of the humanities research process into an actual ontology for use in system design has been conducted by (Benardou et al. 2009) for the DARIAH project and is discussed later in this paper.

The authors concluded their paper with five major implications for cultural heritage digital libraries. First, that the research environment that supports scholarly work is as important as the metadata and searching functionalities. The design and maintenance of such environments they granted, however, will also require a high level of "ongoing technical investment" that is rarely found in the humanities community. Second, they argued that humanities digital libraries will be highly focused and might only include thousands or even hundreds of documents. At the same time, the materials within these collections will have "complex internal structure" and require a large amount of "related contextual material and editorial notes," a feature that is already displayed in many growing digital classics collections. Third, they suggested that in order to become sites that support digital scholarship, humanities digital libraries will need to be created as "bootstrapping tools for their own construction" and support for this will need to be factored in during the design process. Fourth, since projects in the humanities have long life-cyles, both in terms of development and reuse, digital libraries will need to be "developed as an ongoing process with changing audience and needs." Their fifth and final point noted that designing and maintaining such complex libraries also requires high levels of investment.

Creators of Digital Humanities Resources: Factors for Successful Use

Some relevant research reported by the LAIRAH (Log Analysis of Internet Resources in the Arts and Humanities) [486] project has recently confirmed that there have been "no systematic, evidence-based studies of the use and non-use of digital humanities resources"(Warwick et al. 2008b). In order to determine how digital resources were being used or not used, the LAIRAH project utilized log analysis techniques [487] to identify twenty-one popular and well-used digital humanities resources (within the United Kingdom) and then conducted in-depth interviews with their creators to see if common factors that predisposed these resources for use could be identified.

Warwick et al. (2008b) briefly synthesized previous research that had been conducted into scholarly use of digital humanities resource and online information behavior and listed a number of important insights: 1) many scholars were enthusiastic about digital humanities resources but in general preferred "generic information resources" to specialist research sources; 2) humanists needed a wide range information resources and types but their work typically involved reinterpreting "ideas rather than creating or discovering new data or facts"; 3) humanists would only use technology that fit well with their existing research methods and if it saved them time and effort; 4) humanists preferred not to have to take specialized training in order to use a resource; 5) while humanities researchers had "sophisticated information skills and mental models of their physical information environment" they often had difficultly applying these skills in a digital environment; 6) humanities scholars were concerned with the accuracy of the materials they used; 7) scholars wanted information about the analog resource that had been digitized; and finally 8) scholars expected "high quality content" and anything that complicated their use of a resource be it a challenging interface or confusing data, would stop them from using it. These findings, Warwick et al. (2008b) proposed, should be carefully considered by the creators of digital resources:

> Thus it is incumbent on producers of digital resources not only to understand the working practices of the scholars for whom they design, but to produce a resource that is attractive, usable and easy to understand. However, perhaps surprisingly, there appears to be no research that assesses how well digital humanities resources are performing in these respects (Warwick et al. 2008b).

Thus the need to understand the working practices of the scholars for whom a resource is being designed is as important as creating an attractive and usable resource.

While none of the twenty digital humanities projects that were chosen for analysis were within the discipline of classics, the results of the LAIRAH interviews provide some useful information on what makes a digital resource successful in the long-term. Warwick et al. (2008b) explored the documentation (if any) on each website and conducted a semi-structured interview with a project representative that covered the creation and history of a resource, funding, technical standards, dissemination and user testing. Not surprisingly, they found

[486] http://www.ucl.ac.uk/slais/LAIRAH

[487] The LAIRAH project made use of the server logs of the AHDS and the Humbul portal that was merged into Intute (http://www.intute.ac.uk/), a free online directory of academic resources that have been reviewed by subject specialists at seven universities.

that the institutional context and "research culture of particular disciplines" greatly affected the production and use of digital resources. One major issue was limited institutional recognition and prestige for scholars who did digital humanities work, as well as an uncertainty among their colleagues as how to value digital scholarship. Another critical issue for the success of digital humanities projects was adequate technical support and staffing. While most principal investigators (PIs) were relatively happy about the level of support they received (typically from local IT staff or expert colleagues), those that reported contact with a digital humanities center received an even higher level of expert advice. Staffing issues were paramount, as research assistants (RAs) required both subject knowledge and a good grasp of digital techniques. The grant funded nature of most projects also made it hard for RAs to obtain adequate technical training or for PIs to retain them beyond individual projects.

The most important activity, however, that led to resources that were well used was active dissemination of project results. All of the projects that were interviewed spent considerable time disseminating information about their resources at relevant conferences and workshops, and Warwick et al. (2008b) noted that this type of "marketing" was a very new area of activity for many academics. A related if not unexpected finding was that the most well used resources tended to be long-lived. This was not necessarily an indicator of successfully meeting user needs, however according to Warwick et al. " The persistent use of older digital resources, even when newer, perhaps better ones become available," Warwick et al. put forward, "may be explained by a commercial phenomenon known as 'switching costs'"(Warwick et al. 2008b). In other words, users will often remain loyal to a particular resource because the effort involved in switching to a new tool is to great.

Another area that was explored by Warwick et al. (2008b) was the amount of user contact in which successful projects were engaged. They found that few projects had "undertaken any type of user testing" or maintained any formal contact with their users. In addition, most projects had little if any understanding of either how often their resources were used or to what types of use they were put. All projects, however, were interested in how their projects were being used and had made some efforts in this area. The most common method according to Warwick et al. was the idea of "designer as user" or where most PIs assumed that their subject knowledge meant that they understood the needs of users and thus could "infer user requirements from their own behaviour." While Warwick et al. (2008b) granted that some user needs might be discovered in this manner, the only real way to discover user needs they contended is to ask or study the actual users themselves. In addition, they reported that some projects "also discovered that their audience consisted of a much more diverse group of users than the academic subject experts they had expected (Warwick et al. 2008b). A related problem was the lack of non-expert documentation at many projects. In the end only two projects had conducted any type of direct user testing. As with dissemination and marketing, Warwick et al. (2008b) commented that user testing is also not a traditional skill of humanities scholars.

The last major issue determining the success of a digital humanities resource was sustainability. At the time this research was conducted (2007-2008), the AHDS still existed and Warwick et al. stated that many projects were either archived there or backed up on an institutional repository. Despite this archiving, Warwick et al. (2008b) concluded that this older model of final deposit was inadequate since many resources were almost never updated and typically the data was not independent of the interface. In many cases this resulted in digital projects, which despite having large amounts of money invested in their creation, were fundamentally unusable after a few years. This was a problem for which they had few answers, and as they unfortunately concluded "Sustainability remains an intractable problem given the current models of funding and archiving digital resources."

Their final recommendations for the long-term usability of digital resources were for projects to create better documentation, to develop a clear idea of their users and both consult and stay in contact with them, to develop effective technical management and support, to actively disseminate results, and for sustainability, to "maintain and actively update their interface, content and functionality of the resource." All of these recommendations are relevant to the development of any lasting infrastructure for digital classics resources as well.

"Traditional" Academic Use of Digital Humanities Resources

Several studies have recently addressed how "traditional" (e.g. not self-identified e-humanists or digital humanists) academics and students have made use of digital resources and this section will examine the larger findings of this research.

The CSHE Study

One of the largest studies to approach the question of educator use of digital resources in the social sciences and humanities was conducted by the CSHE between 2004 and 2006 (Harley et al. 2006b). This study pursued three parallel research tracks: 1) conducting a literature review and discussions with different stakeholders to map out the types of digital resources available and where the user fit within this universe; 2) discussions (focus groups) with and surveys of faculty at three different types of institutions (within California) as well as with the users of various listservs regarding how and why they used or did not use digital resources; 3) creating a methodology for how user study results might be shared more usefully by interviewing site owners, resource creators, and use researchers. At the same time, Harley et al. argued that the differences between individual disciplines needed to be carefully considered as well as the varying types of users. "The humanities and social sciences are not a monolith, nor are user types," Harley et al. (2006b) explained, "We contend that a disaggregation of users by discipline and institution type allow us to better understand the existing variation in user and non-user behavior." The authors also insisted that there had likely been no "coordinated conversation about user research" across the many types of digital resources available due to the immense variety of such resources found online.

As they began to study the types of digital resources available in order to create a typology of resource types (e.g. data sets, videos, maps, electronic journals, course materials) they quickly discovered the number of resources available was ever growing and that digital resources were being created in different environments by many types of developers. In addition, they also noted that users often defined resources much more granularly than did their creators. The project also defined three major roles for analysis in terms of website "owners" including resource aggregators, developers of tools, and content creators and owners.

The major part of this study consisted of both speaking with and surveying all kinds of faculty in different disciplines [488] and at several kinds of institutions to find out the reasons they either did or did not use digital resources as part of their teaching. Harley et al. (2006b) found that "personal teaching style" and philosophy greatly influenced resource use and that there were a large number of user types from non-users (a diverse group in itself) to novice users to advanced users of digital resources. Images and visual materials were the resources that were listed as being most frequently used, but news websites, video and online reference sources were also used quite heavily. While faculty used Google as their primary means of finding resources, the second most frequently used resource were their own "collections" of digital resources.

The reasons for both use and non-use of digital resources were quite diverse according to Harley et al. (2006b). Major reasons faculty used digital resources included to improve student learning, to integrate primary sources into their teaching, to include materials in teaching that would otherwise be unavailable, and as a means of integrating their research interests into a course. In terms of non-use, the preeminent reason was that the use of digital resources did not support their approach to teaching. Additional reasons included lack of time, resources that were difficult to use, and also notably, the inability to "find, manage, maintain, and reuse them in new contexts." The importance of personal digital collections was also illustrated again, and Harley et al. (2006b) asserted "many faculty want to the ability to build their own collections, which are often composed of a variety of materials, including those that are copyright protected." Thus faculty demonstrated a desire not just for resources that were easier to find and use but also for ones that were "open" in the sense that they could at least be reused in new contexts.

[488] According to the full report, Harley et al. (2006a), 30 faculty from classics participated in this study (11, or 2.4% of the H-Net survey and 19 faculty from California universities (pg. 4-15), and the only other major finding of this report in regards to classicists was that they tended to use fewer digital resources in their teaching than many other disciplines (pg. 4-56)

Interviews with thirteen digital resource providers of "generic" online educational resources (OERs)[489] and two other stakeholders in terms of what types of user research they had engaged in and what they knew about their users revealed that there were no common metrics for measuring use or defining groups of users, but that most projects assumed faculty were their main user group. Their findings largely confirmed those of (Warwick et al. 2008b) that little if any comprehensive or systematic research had occurred:

> The interview analyses suggested that there were no common terms, metrics, methods, or values for defining use or users among the targeted projects. Yet digital resource providers shared the desire to measure how and for what purpose materials were being used once accessed; few providers, if any, however, had concrete plans for undertaking this measurement in a systematic way (Harley et al. 2006b)

Several resource providers were exploring various ways of both engaging with and building a user community as one potential solution to long-term sustainability, a major theme of interviews. "Our research revealed that community building is important to digital resource providers," Harley et al. (2006b) reported, "and many were exploring tools to enable the development or support of user "communities." Some also suggested that community contributions might hold a key to sustainability challenges."

After conducting these interviews, a two-day workshop was held with sixteen experts to discuss OERs and how exploring user behavior might be linked to larger policy or planning issues. Four broad topics were covered by this meeting and included 1) defining a common framework to codify different "categories of content, users, uses and user studies"; 2) the practicality and expense of different types of user studies and methods (e.g. what types of common questions to ask, what level of research must be conducted (formal, informal)); 3) questions of user-demand and long-term-sustainability (curricular, technical/infrastructural, organizational and financial); and 4) the larger research questions that would need to be addressed. The topic of sustainability brought up the largest number of complicated issues. One question that brought up particularly diverse responses in terms of sustainability was whether OER sites should "adapt their content or services to unintended users." "To some participants, unintended use is an opportunity for creative reuse," Harley et al. (2006b) stated, "while many believed that an OER site should not or could not change course to serve an unintended audience." This question was tightly linked with the mission of different OER's and their financial models. In terms of technical/infrastructural sustainability, many participants proposed that OERs and particularly open access ones need a "common place where they can be reliably housed, organized, searched, and preserved" and that "centralized OER repositories" might serve as one answer. Various models for how such a OER repository might be developed were discussed, but a number of participants agreed that federated searching across different repositories would be a "user-friendly" start.

While Harley et al. (2006b) offered a number of conclusions regarding their research findings, a particularly significant one was the great desire of faculty to "build their own re-aggregated resources" or to be able to both use materials from their own personal digital collections and mix them with other digital resources they have found online. The limitations of classroom technologies, the vast array of complicated and typically non-interoperable tools that were available for use in terms of "collecting, developing, managing, and actually using resources," and the inability to integrate many resources with standard learning management systems were all cited as significant challenges. Future digital tool developers, Harley et al. concluded would need to address a number of issues, including the difficulty or "impossibility" of reusing objects that are bundled or "locked" into either static or proprietary resources, complex digital rights issues, uneven interface design and "aesthetics," and growing user demands for resource "granularity" (e.g. being able to find and reuse one image, text, etc. within a larger digital resource).

The LAIRAH Project

While the study conducted by Harley et al. largely focused on how faculty used digital resources in their *teaching*, other related research by the LAIRAH project instead more broadly analyzed academic use of digital resources through the use of "quantitative Deep Log Analysis techniques" and qualitative user workshops. One core goal of their research was to obtain detailed user opinions regarding digital resources and what factors

[489] Resource providers included MIT OpenCourseWare (OCW) (http://ocw.mit.edu/), JSTOR, and the National Science Digital Library (http://nsdl.org/)

inhibited their use (Warwick et al. 2008a). For their log analysis they used logs from the AHDS servers, the Humbul Humanities Hub (now Intute) and Artifact. Unfortunately they were not able to use "individual logs from the servers of digital humanities projects" due to time constraints. In addition to using log data they also mounted a questionnaire and held a workshop with users regarding neglected resources to see if they could determine why resources were not being used. One significant difficulty that they encountered was attempting to extract log data, even when using the logs of large government funded repositories. Nonetheless the log data from the AHDS central site did show those links that were followed on the site and it was thus possible to generate a list of pages that visitors actually used. Resources about warfare were quite popular as were census data and family history. One insight they offered was that resources that were not particularly well named were often seldom used, and they advised digital project creators to utilize simple titles and good descriptions that made it clear what a resource was about.

Since Warwick et al. (2008a) wanted to get a broad range of answers in terms of digital resources, they did not offer a definition of resources in their questionnaire and instead asked participants to list their three favorite resources. Consequently they learned that most of the users they surveyed considered digital resources "not to be specialist research resources for humanities scholarship, but generic information resources." The most popular resource listed was the university library website and this was followed by Google. Many resources were simply classified as "other" and the vast majority were "information resources or gateways, archives and subject portals" as well as subject-based digital libraries. This finding sharply contradicts the belief of many digital project creators that the specialist research tools they create for faculty will be heavily used as Warwick et al. explain:

> It therefore appears that most of our users regard digital resources primarily as a way to access information, which in the analogue world might be compared to the library or archive, rather than specialist research resources which we might compare to a monograph or a literary text for primary study. It is significant that most resources fall into the 'other' category, which suggests that there is a very wide range of resources being used, and very little agreement as to which are most useful (Warwick et al. 2008a)

Similar results were also observed by Dalbello et al. (2006) in that classicists who cited electronic resources never utilized them as research or primary resources, or at least did not admit to doing so.

The last component of the user research conducted by Warwick et al. (2008a) involved a workshop about neglected digital resources. They found it was very difficult to recruit participants and the final group was largely composed of historians, archaeologists, graduate students and individuals who worked in humanities computing. They found that many of their participants, particularly if they came from a more traditional humanities background, were generally unwilling to "commit themselves" in terms of the quality and usefulness of resources, especially if these resources were outside of their discipline. Warwick et al. reasoned that this reluctance was perhaps due to the fact that most "specialist digital humanities research resources" were very unfamiliar to most humanities academics. In general, participants were fairly critical of the resources they were asked to evaluate and there was no "universal enthusiasm" regarding "content, interface and ease of use."

Warwick et al. (2008a) offered a number of general recommendations as a result of this research and strongly argued that publicly funded digital research projects should have to maintain log data and make it available for at least 3 years. They also reiterated that clear and understandable naming and describing of projects was very important for ensuring maximum impact. Additionally since general information sources were largely preferred over specialist research sources, they stated that funding might most usefully be given to projects that create large collections of information resources for reference. As humanities scholars they interviewed demanded the highest possible quality content and interfaces (whether a resource was commercial or free), they suggested that the creators of specialist digital resources should spend more time on interface design and user testing of those designs. While their findings did illustrate that academics both want to easily find and use digital resources, they also articulated that "the kind of scholar who is likely to know they need such a resource and persist until they find it is the kind of early adopter who is already using specialist digital resources"(Warwick et al. 2008a). They thus concluded with a call for the producers of digital resources to focus on drawing in more traditional humanities users.

The RePAH project

More focused research in terms of academic digital resource use has been conducted by the recently concluded RePAH (Research Portals for the Arts and Humanities) project, which carried out a domain-wide survey of how arts and humanities researchers might use an online research portal in their work. A recent article by Brown and Greengrass (2010) has presented an overview of the RePAH project's findings. To set the context for RePAH, Brown and Greengrass first outlined a brief history of e-humanities research, funding and infrastructure within the United Kingdom over the last ten years. Brown and Greengrass highlighted how a major change in strategy had occurred, since an original emphasis on investing in access to resources had shifted to concerns about how these resources were used to questions about the "skill levels and attitudes towards use of ICT in arts and humanities research." At the same time, the authors noted that the arts and humanities involve a large number of disciplines with many different research traditions, and that "it can not be assumed that innovations in one discipline necessarily meet the requirements of others" (Brown and Greengrass 2010).

A research portal the authors explained focused on "federating distributed sites of information" and the RePAH project sought to explore how researchers across disciplines might use such a tool. While Warwick et al. (2008b) earlier reported that there had been no systematic examinations of the scholarly use and non-use of digital resources, Brown and Greengrass confirmed a similar lack of studies involving the general ICT use of arts and humanities scholars as was first noted by Toms and O'Brien:

> Hitherto there has been no sector-wide comparative study to ascertain how researchers are using ICT and what they perceive their future needs to be. Consequently what is needed in terms of an ICT infrastructure to support Arts and Humanities research is not well understood. Are there, for example, significant differences in the ways in which researchers from different disciplines use ICT in their research? Are some domains more technically advanced than others? How widespread is ICT based research across the sector? Can a single portal concept meet the needs of the whole community? (Brown and Greengrass 2010).

Such questions are often difficult to answer, Brown and Greengrass acknowledged, and the fact that responsibility for funding this kind of infrastructure is typically split across multiple agencies makes cross-disciplinary research even more problematic.

In an attempt to answer these questions, the RePAH project broadly defined their set of users as the "arts and humanities research community" and specifically wanted to ascertain information about these users "information discovery strategies," their Internet usage, their awareness and opinions regarding various online services such as repositories and portals, and their future expectations. They utilized a multi-pronged approach that included an online questionnaire (with almost 150 respondents), focus groups, log analysis (the same ones used by the LAIRAH project for their research), Delphi [490] forecasting, and user trials. [491] In the final step of their process, RePAH used the results of the user trials where participants were presented with a range of possible portal features and a number of demonstrators to "cross-check" the earlier results of the focus groups, questionnaires and Delphi exercise.

A number of major findings emerged as a result of this research that illustrate larger trends that need to be considered when designing an infrastructure for any discipline within the arts and humanities such as classics. One major theme identified by Brown and Greengrass they consequently labeled "pull vs. push." While 60% of the respondents to the online questionnaire considered digital resources to be essential to their work, they also saw the Web as a source of data to be used rather than "as a repository into which they could push their own data." While the collection and analysis of information was important to almost half of respondents, data storage and archiving were not given a high priority.

A second major finding that was presented and which had emerged from their focus group discussions was that arts and humanities scholars considered the web to have three major benefits: speed and efficiency, timeliness of resources, and new ways of working. When asked about the shortcomings of the Web, the answers were

[490] Brown and Greengrass explain that Delphi is "a structured process for collecting and distilling knowledge from a group of experts by means of a series of questionnaires interspersed with controlled opinion feedback" and it was used to filter ideas from the focus group by asking experts to rank ideas in terms of their importance to their future research.

[491] According to Brown and Greengrass, "user trials are a technique for gaining user responses to design ideas, working from mock-ups or simulations."

slightly more diverse. While most focus group participants reported being satisfied with the digital resources they had available, they also overwhelmingly wanted greater online access to the subject literature of their field, particularly journals. Participants were also concerned about the large number of low-quality search results they often obtained on the Web, and many wanted "tools for aggregating data for searching and analysis and better quality control and ranking of results"(Brown and Greengrass 2010). Despite wanting better quality control, participants were also suspicious about who would undertake the quality assurance and wanted to have an unmediated role in the process. Other major frustrations with the Web included the lack of interoperability between different digital libraries, increasing access restrictions and intellectual property rights.

The RePAH questionnaire had also focused on the types of resources that scholars used and Brown and Greengrass stated that a wide range of resources were used and there was little commonality between disciplines. In addition, generic sites such as library websites were the most commonly cited, with Google (including Google Scholar, images, etc.), and JSTOR as the next two most frequently cited resources. Interestingly, Brown and Greengrass also commented that in certain disciplines such as classics and ancient history, Google was listed as the "central tool for acquiring digital information," perhaps due to the relatively large number of digital resources in classics. In general, however, the largest category of resource cited was "other."

While user trials elicited some diverse responses to different potential portal features, Brown and Greengrass nevertheless stressed "the overarching message that came out of the user trials was they wanted simple tools that required little or no input of time or personal engagement." Participants highly valued "resource discovery" [492] and filtering tools that provided greater control over web-based resources but also wanted tools that were highly customizable. The most important online resources were journal articles and other bibliographical resources. Workflow management tools such as sophisticated bookmarking features and automated copyright management were also highly desired features. In addition, while participants wanted automatic information harvesting tools to be used against digital content to which they wanted access, the use of these tools against their own "content", however, was considered "problematic." Most collaborative tools such as social bookmarking, collaborative annotation of digital resources, shared document editing, and "contributing to the authentication of digital content" fell in the middle range of desired features. Finally, advanced communication tools (e.g. real time chat and video conferencing) were not considered to be highly valuable and most participants were satisfied with existing systems such as email.

Although Brown and Greengrass believed that the RePAH project had illustrated that digital resources were quite important for arts and humanities researchers, their impact on personal archiving and publishing practices were still very limited:

> …despite its impact on research, ICT has not fed through to the habits and procedures for personal digital data archiving, and has not yet had a substantial impact on the means of scholarly communication in the arts and humanities. In short, it has not yet profoundly influenced the way in which arts and humanities publication is conceived (Brown and Greengrass 2010).

Similar results were observed by Dalbello et al. (2006) and Brown and Greengrass also confirmed the finding of Toms and O'Brien of the picture of the solitary scholar working on a journal article for a printed publication. Additionally, although many scholars wanted greater access to and quality assurance of resources, many were highly distrustful of any portal features that either automatically harvested their content (such as CVs on a research profile page in a portal) or that monitored their activity (e.g. observing the electronic resources they selected from a portal page), even if such features enhanced the performance of the system for the whole community. This hesitance was caused by a number of reasons according to Brown and Greengrass and included the "individualistic nature of the community" and personal privacy fears. Such attitudes and other limited technical understanding, Brown and Greengrass emphasized, however, were important for portal builders to consider in any future design:

[492] This finding is somewhat ironic as the major resource discovery service in the United Kingdom, Intute, has recently announced that its JISC funding will end in July 2011 (http://www.intute.ac.uk/blog/2010/04/12/intute-plans-for-the-future-2010-and-beyond/)

…the concerns raised here suggest a lack of awareness about the extent to which actions are already monitored and recorded. When this is coupled with the strongly expressed preference for simple tools that require little or no learning and their expressions of frustration at the lack of sophistication of search engines (a frustration that was often a function of their lack of familiarity, or perhaps understanding, of Boolean search parameters permitted in Google's advanced search facilities), a picture emerges of researchers with relatively limited technical skills. Our focus group participants reported levels of formal initiation or training in the digital resources that they used varying from little to none. The implication here is clearly that future portal developments should assume only a very basic level of ICT competence (Brown and Greengrass 2010).

The final issue raised by Brown and Greengrass was the importance of *access*, a theme that ran through all of their results, and this access was primarily to online journals. While arts and humanities researchers do desire more sophisticated research infrastructures, Brown and Greengrass concluded, they mostly want open access to content with simple search engines that can nonetheless guarantee quality and relevant results. Portal designers should also assume low levels of IT competence and provide basic interfaces that are also customizable in that the user can add links or feeds to their commonly used sources. They also argued that arts and humanities researchers in general felt no need for tools that support collaborative working, online archiving or electronic publishing, and were unwilling to support tracking systems even if the could support "powerful quality assurance systems."

The TIDSR Study

As has been illustrated by both Brown and Greengrass (2010) and Warwick et al. (2008b), little comprehensive research has systematically explored the impact and actual usage of digital humanities resources. Recent work detailed in Meyer et al. (2009) reports on a study that examined the use of five digitization projects within the United Kingdom using a variety of measures in order to obtain a more nuanced picture not just of these projects but of "digitized material in general." This JISC funded research was undertaken to promote standards and knowledge sharing among different projects in terms of how to measure the usage and impact of their resource. One major outcome of their work was the creation of a "Toolkit for the Impact of Digitised Scholarly Resources" (TIDSR) [493] While Meyer et al. acknowledged much was learned from both the LAIRAH and CSHE studies, they also suggested that both these studies were missing one part of the larger picture:

One major missing part, however, is any concrete way for collection managers, developers, and funding bodies to attempt to understand and collect data for measuring impact from the onset of a project and throughout the life-cycle of a digitisation effort. The toolkit is an attempt to fill this gap (Meyer et al. 2009).

For this study in particular they chose five digital projects [494] and used a variety of methods including quantitative measures (webometrics, bibliometrics, log file analysis) and qualitative methods (content analysis, focus groups and interviews). [495] A number of important themes emerged from the interviews they conducted with project creators including the importance of close contact with users when developing a project, how many digitized projects had either little or almost no contact with the "custodians of the original content" that had been digitized, and interestingly a "discrepancy between intended usage and perceived success" or in other words that many project creators discovered that the uses to which their collections had been put were very different than how they thought they would be used. Interviews with users were used to gauge the varying levels of project impact and they noted several trends including that the quality of some undergraduate dissertation work seemed to improve through contact with primary sources, that some new types of research were being presented at conferences (e.g. increasingly quantitative research was found in many conference papers for the fields served by the relevant digital humanities resources), and that some new types of research were also being attempted. At the same time quantitative research projects with digital data also had some problems, such as those reported by one researcher who stated that keyword searching was still very unreliable for digital newspaper data where

[493] http://microsites.oii.ox.ac.uk/tidsr/

[494] The five resources were Histpop(http://www.histpop.org), 19th Century British Newspapers (http://www.bl.uk/reshelp/findhelprestype/news/newspdigproj/database/paperdigit.html), Archival Sound Records at the British Library(http://sounds.bl.uk), BOPCRIS (http://www.bl.uk/reshelp/findhelprestype/news/newspdigproj/database/paperdigit.html), and Medical Journals Backfiles (http://library.wellcome.ac.uk/backfiles).

[495] While the research presented in Meyer et al. 2009 focuses on the results of the qualitative measures, the full report can be viewed at : http://microsites.oii.ox.ac.uk/tidsr/system/files/TIDSR_FinalReport_20July2009.pdf

the text had been created by OCR. While there were still far too many false negatives and positives, the researchers still observed that searching the digitized newspapers was still far superior to using microfilm.

In addition to these individual interviews, Meyer et al. also held focus groups with two groups of students and these revealed that in general the students were enthusiastic about digital resources and that undergraduate students used digitized collections on a regular basis, both ones recommended by their tutors and ones that they sought out independently. Somewhat different results were reported by postgraduate and postdoctoral students who were also using resources that had been recommended but were uncertain about where the best recommendations would come from and were also far more skeptical about the quality of resources that they discovered outside of library catalogues and finding aids. [496] One behavior reported by both groups of students, however, was a general unwillingness to cite digital material [497] that they had used:

> Both groups were unlikely to cite the digital material if there was a paper or analogue citation available, although for different reasons. The undergraduates were concerned that they would be perceived as having not completed 'proper' research unless they cited the analogue resources, whereas the postgraduates and postdoctoral researchers were" more concerned about giving stable citations that future researchers would be able to trace (Meyer et al. 2009)

The other major reasons students gave for not using digital resources were trustworthiness, the persistence of a digital resource, general vetting concerns the and importance of branding by trusted institutions to promote use of a digital resource. The concern for needing stable citations to electronic resources was also illustrated by Bodard (2008) in his discussion of the creation of the *Inscriptions of Aphrodisias* website.

In order to better determine the actual impact and use of a digital resource, Meyer et al. had a number of suggestions. To begin with they argued that digital projects should plan on measuring impact from the very beginning of the project, ideally before the website has even been designed. While they also believed that impact should be measured regularly, they also advised that projects should not get "bogged down" by planning overly detailed studies. In addition, they proposed that sustainability strategies should be built in from the beginning. Other recommendations included that projects should make efforts to secure follow up funding to measure impact and also actively promote their project through blogs, publications, conference reports, etc. as well as make sure that they are included in trusted gateways (such as library information portals). In the long run in terms of measuring impact they urged that all projects should consider multiple sources of evidence, examine them from different perspectives and use a variety of metrics. On a practical note they pointed out that projects that don't maintain stable and easy to cite URLs make it difficult for scholars to reference them in their publications. Lastly, they recommended reaching out to the next generation of scholars. "There are important generational shifts taking place: younger researchers are developing research habits that will become mainstream as they replace their elders." Meyer et al concluded, "These so-called digital natives are a natural constituency for digital collections, so ensure that your resources are available to them, usable by them, and promoted to them" (Meyer et al. 2009).

Overview of Digital Classics Cyberinfrastructure

Requirements of Cyberinfrastructure for Classics

A number of recent research studies have explored some of the potential needs of a cyberinfrastructure for classics, including the development of working paper repositories, the creation of new collaborative models for

[496] This behavior is in contrast to the digital humanist researchers observed in Toms and O'Brien who relied largely on Google or other search engines to find resources of interest, but supports the findings of (Warwick et al. 2008a) and (Brown and Greengrass 2009) that academics and students valued resource discovery tools that helped them identify reliable digital resources. In contrast, a recent survey of over 3000 faculty by the Ithaka group indicated that faculty were increasingly using discovery tools other than "library specific starting points" and that only 30% of humanities faculty still started their search for digital materials using a library discovery tool (Schonfeld and Housewright 2009).

[497] Similar citation behavior was reported by (Sukovic 2009) where literary scholars and historians were often unwilling to cite digital resources (and thus often cited the analogue source even when they had only used the digital version) for various reasons, including fear that their colleagues did not approve of using such resources and that the referencing of digital resources did not fit within the academic practice of their discipline. She concluded that "The multifarious nature of scholars' use of e-texts, revealed in the study, was not reflected in citation practices."

scholarship and teaching, the requirement of open data and collections, and the large variety of services that will be necessary.

Open Access Repositories of Secondary Scholarship

Two recent studies have focused on the potential of open access repositories for classical studies (Ober et al. 2007, Pritchard 2008). Ober et al. discussed the creation of the open access working papers repository, the Princeton Stanford Working Papers in Classics (PSWPC), [498] and also examined the potential benefits of electronic publishing and the relationship of "working papers" to traditional publishing. The PSWPC is a web-based repository that is open to the faculty and graduate students of Stanford and Princeton and the papers are not formally peer reviewed. Nonetheless many contributors have put up preprints or working papers that eventually went on to be formally published. The creation of this repository has raised a number of issues regarding long-term access and preservation, which might be better guaranteed by a commercial archive. The authors define three processes as the traditional roles of scholarly publishing: making research public, certification and archiving, and also propose that the process of certification or peer review is the most important role of traditional publishing.

While the authors acknowledge that the only assurance of value of the working papers in the PSWPC is the academic standing of the two departments, they also point out that a large amount of traditional publisher peer review is relatively undemanding. [499] They also remark that a distinction needs to be made between "preprint/working paper" archiving and post-print archiving, or the archiving of a paper that has already been formally published. One disappointment they also noted was that neither the American Philological Association (APA) nor the Archaeological Institute of America (AIA) had yet created large working paper repositories for the entire disciplines. Ober et al. also offer a number of recommendations for humanities scholars in working towards open access: 1) promotion of pre-print and post-print archiving to the largest extent possible 2) try to get the larger professional organizations involved 3) that academic authors fight harder to retain their copyrights and 4) that all institutions in higher education should "move to greater flexibility in considering what counts as 'publication' in the new electronic media" (Ober et al. 2007).

A recent paper by David Pritchard provided an external look at the PSWPC and briefly explores the large issues of open access, cyberinfrastructure and classics. The PSWPC had been far more successful than it anticipated, reporting almost 2000 downloads a week in September 2007. Pritchard suggested that the PSWPC fulfilled two important scholarly tasks: 1) making a far greater wealth of classical scholarship available to a wider audience and 2) helping the authors solicit feedback and find a greater audience for their work. He also proposed that there were four reasons for the success of the PSWPC, first that it allowed specialists to share research, second was the already "entrenched use of computers by ancient historians and classicists," a phenomenon he noted surprised many non-classicists, third, the demand for open access research in general, and fourth, the high quality of the papers. Pritchard did suggest, however, that the PSWPC might improve by including better metadata and make the repository harvestable through OAI-PMH. Like Ober et al, Pritchard recommended that authors should seek to archive their pre or post-prints, but his views on cyberinfrastructure were limited to the creation of institutional repositories or more departmental collections of work papers.

Open Access, Collaboration, Reuse and Digital Classics

In addition to open access repositories of scholarly publications, [500] the need for greater openness in terms of data, collections, methodologies and tools and the new models of collaborative scholarship such openness might

[498] http://www.princeton.edu/~pswpc/index.html

[499] A recent blog post by Kent Anderson at "The Scholarly Kitchen" has provided an interesting look at the varying processes and quality of peer review by different commercial publishers (http://scholarlykitchen.sspnet.org/2010/03/30/improving-peer-review-lets-provide-an-ingredients-list-for-our-readers/). In addition, a recent article by (Bankier and Perciali 2008) has suggested that the creation of peer-reviewed open-access journals may help to revitalize digital repositories and provide a natural publishing outlet for universities.

[500] While the largest number of open access publications of classical scholarship are typically reviews, working papers and journal articles, several scholars have made copies of books available, including Gregory Crane, *Thucydides and the Ancient Simplicity: The Limits of Political Realism*, http://www.escholarship.org/editions/view?docId=ft767nb497&brand=ucpress and Gregory Nagy's *Pindar's Homer: The Lyric Possession of an Epic Past*. http://www.press.jhu.edu/books/nagy/PH.html

support have received growing attention. As this review has illustrated a number of projects have made their texts and source code openly available such as the archaeological projects found in Open Context, the Perseus Digital Library, Pleaides, and the Inscriptions of Aphrodisias. Similarly, many authors (Crane, Seales and Terras 2009, Bagnall 2010, Bodard and Garcés 2008, Bodard 2009, Robinson 2009) have called for a new level of open access that not only provides *access* to scholarship and collections but also provides and promotes *openness*, in terms of actively supporting the *reuse* of source code, data and texts.

Reuse is not always an easy task, however, as evidenced by both the Hestia and LaQuAT projects. Furthermore, documenting reuse is also often difficult as Gabriel Bodard explained in his introduction to a 2009 Digital Humanities panel on the reuse of open source materials in ancient studies, for "there is often relatively little evidentiary support in the form of openly published datasets that have been independently tested or re-used by other projects" (Bodard 2009). This panel included the LaQuAT project, the Homer Multitext and Pleiades, and Bodard listed several important insights including the need for open licensing in addition to making materials available online, the conflict between electronic publication as "resource creation" vs. "self-contained research output," and the need to convince scholars of the advantages of publishing source code and methodologies and not just polished conclusions. Bodard also complicated the idea of reuse by arguing that digital projects need to consider how to support more sophisticated reuse strategies, including the possibility that materials will be reused in unexpected ways and determining how to enable not just improved access or a better interface to a collection but to actually allow the creation of new interpretations or aggregations of data.

Neel Smith has also echoed this point and explained that the Homer Multitext project has chosen to use open data formats, well defined standards and has released all software under a GNU public license not just to ensure sustainability and digital preservation but also to promote the highest amount of reuse of the material as possible. "From the outset, the Homer Multitext Project has been shaped by a sense of our generations' responsibility, as we transform the Iliadic tradition into yet another medium, to perpetuate as completely as we can the tradition we have received," Smith articulated, "We need to ensure that as we focus on the new possibilities of digital media we do not inadvertently restrict what future scholars and lovers of the *Iliad* can do with our digital material" (Smith 2010, pg. 122). In terms of freedom and reusability, Smith recasts Richard Stallman's four kinds of freedom for free software, [501] or the freedom to run, study, redistribute, and improve and release in terms of the Homer Multitext. The freedom to run includes the ability to read a text or view an image, to study includes the ability to see how resources are encoded, to redistribute involves the ability to share and redistribute the digital objects, and to improve & release includes the ability to edit and resample or redistribute texts and images.

In addition, the abstract object model of the Homer Multitext has been translated into an application architecture that will ensure that the "*functionality* of Multitext applications can persist as easily as the data in our simple archival storage formats" (Smith 2010, pg. 132). This has led the Homer Multitext Project to adopt four basic architectural principles: 1) in order to support reuse of code APIs were used for "distinct components of the system"; 2) "independent decoupled components" were used whenever possible; 3) all these components have been exposed to the Internet; 4) all software has been released under a GNU license. As Smith succinctly concludes: "Taken together, these principles lead us to an architecture built on *a suite of self-contained network services with explicit APIs, implemented in free software"* (Smith 2010, pg. 132).

For many of the earliest digital classics projects, however, the primary concern was open access, a revolutionary move in itself at the time, and this was primarily defined as providing free access to scholarship on the web. One of the earliest projects to follow this model was the *Bryn Mawr Classical Review* (BMCR). [502] According to its website, BMCR "publishes timely reviews of current scholarly work in the field of classical studies (including archaeology)." BMCR began as a listserv in 1990, the first gopher site became available in 1992 and the current website emerged from a partnership with the Stoa Consortium. The entire archive of BMCR reviews (from 1990 onwards) is available on this website and can be browsed by year, reviewer, or author of work.

[501] http://www.gnu.org/philosophy/free-sw.html
[502] http://bmcr.brynmawr.edu/

There are a large number of reviewers that participate in the BMCR and all reviews have stable URL's so that they can be cited and linked to easily. Since August 2008, the BMCR has also offered a blog [503] that publishes citation details and a link to reviews on the BMCR website as individual blog entries, so that users can both subscribe to the blog and get updates of BMCR content through a blog reader program as well as leave comments on reviews. In addition, the BMCR also provides a daily email digest as another way of pushing out its content.

A larger undertaking that also focused on creating a collaborative community and new digital publishing opportunities is the "Stoa Consortium for Electronic Publication in the Humanities," [504] often simply referred to as Stoa. This consortium was founded in 1997 by the late Ross Scaife and according to its website exists to serve a number of purposes: "dissemination of news and announcements, mainly via the gateway blog; discussion of best practices via discussion groups and white papers; and publication of experimental on-line projects, many of them subject to scholarly peer review." In addition, the Stoa consortium firmly states that "open access to networked scholarship" is one of their strongest principles. This is strongly illustrated by the large number of hosted projects at Stoa, including: Ancient City of Athens (a photographic archive of archaeological and architectural remains of ancient Athens intended for students and teachers), [505] *Ancient Journeys* (an online festschrift), [506] *Confessions* [507] of Augustine (an online reprint of the text with a commentary by James J O'Donnell), Demos (a growing digital encyclopedia about Athenian democracy extensively cross referenced with Perseus), [508] Diotima (an interdisciplinary resource on women and gender in the ancient world), [509] Metis (a repository of QuickTime movies of Greek archaeological sites), [510] and Suda Online (a collaborative online version of the Suda). [511] The website also notes that many projects at the Stoa are closely linked with materials and tools from Perseus and it is also closely affiliated with the Digital Classicist website and community.

The Digital Classicist [512] website has been established as a "decentralised and international community of scholars and students interested in the application of innovative digital methods and technologies to research on the ancient world." The site is not officially hosted by any institution but it serves as a web-based hub for communication and collaboration among digital classicists. Every summer the Digital Classicist hosts a series of seminars [513] at the Institute of Classical Studies in London where practitioners can present cutting edge research on the use of computational methods in the study of antiquity. The largest component of the Digital classicist website is the wiki, however, which was first created by Gabriel Bodard and other practitioners who were interested in the "application of the digital humanities to the study of the ancient world" (Mahony 2006). This site aimed from the beginning to bring scholars together to support collaborative working and thus formed partnerships with the Stoa Consortium, the CHS, and the Digital Medievalist blog. [514]

The Digital Classicist wiki [515] was thus created as a central location to link together the diverse scholarship in the various areas of ancient studies, and even more importantly, sought to "fill an important gap in the existing scholarly documentation by creating concise, reliable and critical guidance on crucial technical issues for scholars who may only be interested in a basic introduction to such issues with links to further resources if they wish" (Mahony 2006). The website, however, also meets two other important needs of digital classicists, who according to Mahony and Bodard, require a space for both building *communities* and working *collaboratively*:

---

[503] http://www.bmcreview.org/
[504] http://www.stoa.org/
[505] http://www.stoa.org/athens/
[506] http://www.stoa.org/lane/
[507] http://www.stoa.org/hippo/
[508] http://www.stoa.org/projects/demos/home, and for more on Demos, see (Crane et al. 2006).
[509] http://www.stoa.org/diotima/
[510] http://www.stoa.org/metis/
[511] http://www.stoa.org/sol/
[512] http://www.digitalclassicist.org/
[513] http://www.digitalclassicist.org/wip/index.html
[514] http://www.digitalmedievalist.org/
[515] http://wiki.digitalclassicist.org/Main_Page

The most striking and successful aspect of Digital Classics is its sense of community and collaboration. Digital Classicists do not work in isolation; they develop projects in tandem with colleagues in other humanities disciplines or with experts in technical fields…They do not publish expensive monographs destined to be checked out of libraries once every few years; they collect data, conduct research, develop tools and resources, and importantly make them available electronically, often under free and open license such as Creative Commons, for reference and re-use by scholars, students, and non-specialists alike (Mahony and Bodard 2010, Introduction, pg 2).

Anyone can join the Digital Classicist wiki by simply applying to one of the four editors for a account. One major component of the wiki is a directory of over 90 digital classics projects organized alphabetically. The length of project descriptions can vary and not all descriptions are linked to active websites. Additionally, a FAQ has a list of 45 articles on best practices in digital classics and includes diverse topics from "Concording Greek and Latin Texts" to "Sanskrit, typing and display." The website also includes a list of 33 tools from "advanced imaging techniques" to "TimeMap" and a brief list of selected electronic resources is also included. This wiki provides an excellent means of entry for scholars first exploring potential the application of digital technology in their area of interest and also provides many collaborative working opportunities.

Although collaboration is a frequently lauded virtue of many digital projects such as Stoa and the Digital Classicists, one scholar quoted by Harley et al. (2010) stated rather bluntly that the level of collaboration could vary in classics depending on the discipline:

I would say collaboration is still relatively rare in the literary side of the classics. Not that many people will coauthor articles on Socrates…This may be different for projects with more technical components, like archaeology, papyrology, or epigraphy…In those areas, there are a lot of projects that require collaboration…I would say that those particular fields—epigraphy, which is reading rock inscriptions, and papyrology, working with bits of papyri—are enormously collaborative…I also think classics, on the whole, has not done too badly in embracing other fields…or at least certain practitioners of classics have gone out there and hooked up with colleagues in various disciplines and brought things back that have continued to expand the field or expand the range of things we can do (Harley et al. 2010, pg. 102).

Harley et al. also noted that while may scholars often worked independently in terms of the close study of documentary remains, they often frequently worked together in the creation of scholarly editions, exhibition and digital projects. On the other hand, the desire for collaborative work, even with documentary remains has been illustrated by the VRE-SDM.

One of the largest and oldest truly collaborative digital classics project is the Suda On Line (SOL), a "massive 10th century Byzantine Greek historical encyclopedia of the ancient Mediterranean world, derived from the scholia to critical editions of canonical works and from compilations by yet earlier authors." [516] The purpose of SOL is to create a keyword searchable and freely available XML encoded database of this encyclopedia complete with translations, annotations and bibliography as well as automatically generated links to other electronic resources. Over 170 scholars from 18 countries have contributed to this project, and 25,000 of the 30,000 entries have been translated. As explained by Anne Mahoney (2009), this collaborative translation project has made the Suda text available to non-specialists and the on-line edition is far easier to use than the print. "As a collaboration," Mahoney declared, "SOL demonstrates open peer review and the feasibility of a large, but closely focused, humanities project" (Mahoney 2009).

In her brief history of the SOL, Mahoney reported that it was one of the first collaborative encyclopedias and predated Wikipedia by several years. Many of the original encyclopedia entries in this unique reference work were also filled with incorrect information, so each digital entry contains explanatory commentary and references. The SOL also serves as an important source of both fragmentary texts and text variants. "Its authors had access to some texts that are no longer extant, so there is material in the Suda that cannot be found anywhere else," Mahoney noted, "They also had different editions of some of the texts we still read, so quotations in the Suda may reflect variants that are not preserved in our textual tradition" (Mahoney 2009).

The SOL was implemented online as a semi-structured text and the translation and editing of the encyclopedia are still ongoing. Prospective translators have to register and then ask to be assigned specific entries. While

[516] Reference works seem to lend themselves to collaboration, for examples consider "DIR: De Imperatoribus Romanis: An Online Encyclopedia of Roman Rulers and Their Families" (http://www.roman-emperors.org/), a collaborative encyclopedia that includes Roman and Byzantine biographies prepared by scholars and actively updated and linked to other classics sites, and Vicipaedia, a Latin Wikipedia (http://la.wikipedia.org/wiki/Pagina_prima).

there are many translators that work on this project, only a subset are also designated editors that also have the authority to change translations. All editors have significant ability in Ancient Greek and many are college and university professors. Some of the primary goals of editors are to augment bibliographies, add commentaries, and verify that translations are correct for SOL entries. The editorial mechanisms of SOL also serve, according to Mahoney, as a "type of peer review process." Each entry credits its original translator but also the editors who have worked on it and this process allows the recognition of all scholars involved and serves as a clear contrast Mahoney notes to the blind reviewing found in many classics journals. The most critical point of this process, however, Mahoney asserted is that it demonstrates the *ongoing* nature of scholarship:

> Perhaps more important, SOL shows how scholarship progresses. A translation or commentary published in a book appears final and finished; readers are not given any clues about how it came into being. SOL's translations and commentaries show the process of successive refinements, demonstrating that first drafts are almost never perfect, and that even senior scholars' work can benefit from editorial attention (Mahoney 2009).

Interestingly, Arne Flaten (2009) made similar arguments regarding how the creation of digital architectural models in the Ashes2Art project that represented uncertainty and various scholarly interpretations illustrated to students the ongoing nature of scholarly arguments.

Mahoney also pointed out that the SOL demonstrates how the digital environment often provides a far more natural way to exploit the knowledge found within a complicated reference work. While the SOL is not a completely new work, it is also not simply a digital reproduction of the printed one. The environment of the web makes it possible to better illustrate the "commentary nature of the Suda" as Mahoney details, because quotations can be identified and labeled, explicit references to primary source texts can be hyperlinked to, and bibliographies can be expanded to include modern relevant scholarship. At the same time, translators can also add links to any online resources they find useful, including ones far beyond the traditional bounds of classical scholarship. Ultimately, Mahoney concluded that the most important accomplishment of SOL was that this material was now available to a far wider audience. Expanding the opportunities of collaboration beyond scholars to the interested public was considered as important by a variety of projects.

While this section has largely focused on access in terms of digital scholarship and content that is freely available, another key component of access is the ability to *find* such materials in the first place. The nature of open access digital collections in classics and the challenges of both cataloging and collecting them has been addressed by Chuck Jones, director of the library at ISAW. [517] As his charge at ISAW is to "develop a library of the scholarly resources required to support a research and teaching program covering the ancient world from the Pillars of Hercules to the Pacific and from the emergence of civilized life until Late Antiquity," he quickly realized that such a collection would have to be both physical and digital and that ultimately the digital component of the ISAW library would include resources both developed locally and elsewhere (Jones 2010). The ISAW is also seeking to develop a project they are calling the "Ancient World Digital Library" to integrate points of access and discovery to materials within tools scholars already use.

As the chief editor of the Abzu [518] bibliography (first started in 1994 and now part of ETANA), Jones described the changing nature of his cataloging work from almost anything he could find to conscious collection making. Whereas once he focused on also including access to commercially licensed materials, he found that research library finding tools covered this area well. At the same time, however, he realized: "It was equally evident that the research library community was not yet coming to grips with providing suitable access to born-digital and open access digital publication which is freely distributed, requiring neither purchase nor license"(Jones 2010). So as the work of Abzu continued, Jones also decided to create the blog "the Ancient World Online" [519] as a means of providing even faster access to new open access publications on the Ancient World. While Jones had originally blogged solely under the larger Ancient World Bloggers Group, [520] he found that the sheer volume of resources available online necessitated the development of his own blog specifically dedicated to open access

[517] http://www.nyu.edu/isaw/
[518] http://www.etana.org/abzu/
[519] http://ancientworldonline.blogspot.com/
[520] http://ancientworldbloggers.blogspot.com/

sources about the Ancient World. For example, Jones maintains an alphabetical list [521] of open access journals in Ancient Studies that is continuously growing and currently has over 600 titles. [522] This list also demonstrates that the idea of providing open access to scholarship is steadily gaining acceptance within the classical community.

A detailed analysis of the history and results of one of these open access journals (Frankfurter elektronische Rundschau zur Altertumskunde (*FeRA*) has been explored in a recent Archaeolog blog post by Stefan Krmnicek and Peter Probst (Krmnicek and Probst 2010). [523] They explained that *FeRA* was created in 2006 and was intended as an online forum for young scholars in archaeology from all over the world to publish their work. *FeRA* is published three times a year and has included 36 contributions in German, English and Italian. In analyzing their log files, they noted that only about 14% of their visits originated from academic networks, and while they acknowledged that many academics might utilize commercial ISPs to access *FeRA*, they believed these results also suggested that "a fairly large group of people interested in the very specialized field of classical studies exists outside academia." At the same time, they also revealed that the number of manuscripts submitted by young scholars had been far less then expected and that the emphasis had shifted from articles to reviews, and they hypothesized that scholars that were not yet established in their fields were reluctant to publish outside traditional print media, a supposition confirmed by the research of Harley et al. (2010). Thus the challenges of traditional peer review and scholarly promotion meant that fewer younger scholars were fully profiting from opportunities in digital publishing (e.g. reaching a greater audience, higher research impact).

Finally, there are a number of prominent blogs that explore scholarship on the ancient world. As listed above, the Ancient World Bloggers Group is a meta-blog with many bloggers and it serves as "a place for posts and discussion about blogging the Ancient World." Two other prominent classical blogs are antiquist and [524] RogueClassicism. [525] While a full review of blogs is beyond the scope of this review, Tom Elliott has put together several feed aggregators [526] that bring together a large number of blogs including "Maia Atlantis: Ancient World Bloggers" that brings content together from bloggers at the Ancient World Bloggers Group and the eClassics community on Ning [527] and "Electra Atlantis: Approaches to Antiquity" that brings together content from ancient world blogs that also frequently examine issues of digital scholarship and technology. These aggregators are excellent tools for keeping current on the classical blogosphere.

Undergraduate Research, Teaching and E-Learning

While the previous section discussed new forms of openness and collaboration among scholars, the field of digital classics has also presented new opportunities for collaboration with students through undergraduate research. In addition, the large number of digital classics resources online as well as the number of websites designed for independent learning present new possibilities for teaching. This section will look at some recent efforts in these areas.

There are a number of useful e-learning resources online for both traditional students and independent learners of classical languages as well as for those studying the ancient world. [528] One of the oldest resources available is Textkit, [529]a website that provides a number of free online resources for the learning of Greek and Latin. Some

[521] http://ancientworldonline.blogspot.com/2009/10/alphabetical-list-of-open-access.html
[522] This list also illustrates the importance of providing such a collection service for when searching in the Directory of Open Access Journals (DOAJ) (http://www.doaj.org/), various keyword searches (ancient (6 journals), antiquity (9 journals), classics (6), classical (14)) turned up only 24 unique classics journals, including three of the most prominent, *Didaskalia* (http://www.didaskalia.net/journal.html), *Electronic Antiquity* (http://scholar.lib.vt.edu/ejournals/ElAnt/) and *Leeds International Classical Studies* (http://www.leeds.ac.uk/classics/lics/)
[523] http://traumwerk.stanford.edu/archaeolog/2010/05/open_access_classical_studies.html
[524] http://www.antiquist.org/blog/?page_id=2
[525] http://rogueclassicism.com/
[526] http://planet.atlantides.org/
[527] http://eclassics.ning.com/
[528] Many thematic resources have been developed for the study of particular aspects of classics online. For example, the study of ancient medicine includes "Medicine Antiqua" (http://www.ucl.ac.uk/~ucgajpd/medicina%20antiqua/index.html) a selected classical text repository and online resource directory created by the Wellcome Trust Centre for the History of Medicine at UCL and Asclepion (http://www.indiana.edu/~ancmed/intro.HTM) "a World Wide Web page devoted to the study of ancient medicine" that was created by the University of Indiana Bloomington.
[529] http://www.textkit.com

of its core collections are a number of public domain grammar books as well as a number of Greek and Latin e-texts. Textkit also has an extensive forum where after registering users can participate in various topics about learning Latin and Greek.

Another long-standing project is VRoma, [530] an online learning community of teachers and students that is dedicated to creating online resources for "teaching about the Latin language and ancient Roman culture." This project was initially funded in 1997 through a "Teaching with Technology" grant from the NEH but still maintains an active website that has two main components: an online learning environment (MOO) and a collection of Internet resources. This MOO simulates an online "place" or "a virtual learning environment that is built upon a spatial and cultural metaphor of ancient Rome." To explore this virtual space users can either log in as guests or can apply for a VRoma character and password.

Another online learning environment is Silver Muse, [531]a resource that was created by the Classics Department of the University of Texas-Austin that seeks "to provide a Web-based system to teach and promote research in Latin epic poetry of the early empire" and includes authors such as Ovid and Lucan. The Silver Muse system provides a hypertextual reading environment of the text of the poets, with linked reading guides, commentaries and essays. The reader can access the full text of a number of works and click on any word to get both a translation and an example sentence.

The Alpheios Project, [532] which makes software freely available for reading and learning languages, has recently released several tools for reading Greek and Latin online. [533] These tools are Firefox extensions that add some specific functionalities to the browser and are usable with any HTML and Unicode compliant texts. After downloading the Alpheios toolbar, the user must first choose either Greek or Latin, and can then utilize several important features that include looking up a word by either clicking on it or by entering it in the toolbar, listening to how a word is pronounced, [534] and a personal vocabulary tool that stores the words you have looked up. The Alpheios website also provides access to a number of "Alpheios Enhanced Texts" and when reading these texts the toolbar has an additional feature that allows the user to access diagrams of each sentence in the form of a dependency tree. Users can also create and save their own dependency tree diagrams of sentences.

The above list of resources is just a small sample of the wealth of material online for both the formal and informal student and opportunities for collaboration were particularly evident in TextKit and VRoma. The possibility of meaningful participation by students in more formal classical teaching and scholarship is a more difficult proposition, but one that Blackwell and Martin (2009) counseled can be addressed by new models of undergraduate research. While the potential of undergraduate research was also considered earlier in the discussion of the Ashes2Art project (Arne 2009) and through the creation of scholarly treebanks (Bamman, Mambrini and Crane 2009), Blackwell and Martin examine the potential of several digital classics projects, particularly the Homer Multitext (HMT), to provide students with new research and publication opportunities. The traditional task of teaching students to read scholarship and produce essays with primary and secondary source citations, Blackwell and Martin argued, needs to be revamped for the digital world. One way to engage undergraduates in scholarship they suggested was to have them create online publications that would thus be read by more than just their teacher and that made extensive use of actual primary sources (rather than relying solely on secondary sources that reference them).

---

[530] http://www.vroma.org/
[531] http://uts.cc.utexas.edu/~silver/
[532] http://alpheios.net/content/alpheios-texts
[533] The Alpheios project has made extensive use of various resources of the Perseus Digital Library, including both the Ancient Greek and Latin treebanks. The code for their tools can be downloaded from (http://sourceforge.net/projects/alpheios/)
[534] This feature utilizes the open source tool eSpeak Speech Synthesizer (http://espeak.sourceforge.net/). Another online learning resource that includes audio samples of Greek and Latin is the Classical Language Instruction Project at Princeton University (http://www.princeton.edu/~clip/). This website contains samples of scholars reading Greek and Latin prose and poetry in order to help students get acquainted with the sounds of Greek and Latin and to practice their reading skills. The authors include Homer, Plato, Pindar, Virgil and Seneca. Another unique audio resource is "Ancient Greek Music" http://www.oeaw.ac.at/kal/agm/index.htm, a website that contains recordings of "all published fragments of Ancient Greek music which consist of more than a few scattered notes"

A related challenge of this proposal, however, is the need for far more access to both primary and secondary materials online so students can both make use of them and link to them. While the authors granted their views of the potential for undergraduate scholarship once "all the sources are online" might be somewhat idealistic, they still held high hopes:

> The very effort of examining primary sources and thinking about their possible meanings would bring home the reality that scholarship is always research, in the sense of finding, identifying, interpreting, and presenting evidence. Students could operate as scholars, whether through the process of verifying the plausibility of the presentation of evidence by others, or by presenting arguments and interpretations that are in one way or another original, in all the various senses of that word…When all the sources are online, then we as teachers of Classics can more effectively engage our undergraduate students as collaborators in research, whether in the collection of, for example, themed primary source collections, or in the interpretation of the countless issues in Classics and ancient history that still await effective investigation based on careful analysis of well-chosen and clearly defined data sets rather than impressionistic assertions (Blackwell and Martin 2009).

One salutary effect of having all primary sources online, Blackwell and Martin articulated, would be that more scholars might feel obligated to be far more meticulous about their own standards of primary source citation. As an example, they mentioned the confusing scholarly practice of citing quotes of fragmentary authors by the standard reference system of a particular edition of collected fragments without also citing the primary text from which the fragmentary quotes were originally drawn, a practice that makes it very difficult for students to decipher these references. Another important method for undergraduates to contribute to classical scholarship Blackwell and Martin offered was through the creation of lists and indices. They noted that as more resources became available online through open access publication and as more software tools were able to aggregate data from wide-ranging sources, the creation of lists and indices would become far more important.

The most significant opportunity, however, had come through the HMT project, where starting in 2006, a grant was secured by Casey Dué of the University of Houston to pay for undergraduate research assistants at this university as well as the College of the Holy Cross and Furman University to begin working on the project. This group of undergraduates called the HMT fellows were given the task of creating XML transcripts and translating specific texts of 5 Byzantine and medieval manuscripts of the *Iliad*. One important research question being considered by the HMT editors was how the editions of Aristarchus differed from the medieval editions and how a drift in the language might indicate "the notion of an ongoing tradition of multiformity." [535] While traditional critical editions of Homer typically obscure these differences, the HMT editors hoped that the work of the HMT fellows in creating XML transcripts would "highlight a problem in the history of the Homeric text, thus contributing a point of conversation and analysis to the ongoing study of the Iliad"(Blackwell and Martin 2009). Blackwell and Martin submitted that this new collaborative model of research, which produced "electronic texts (not required to be "printable") in transcription (rather than collation)," not only allowed both students and professors in a distributed geographic environment to work with high quality images of "primary texts, the papyri, the Byzantine and medieval manuscripts" but also supported a new type of scholarship that addressed the limitations of traditional critical editions of Homer.

Ultimately Blackwell and Martin concluded that the integration of both information technology and new models of faculty-student collaboration into conventional classical teaching would be necessary not just to reinvigorate the discipline but also to keep it relevant:

> Because technology has lowered the economic barriers to academic publishing—a reality that too few publishing Classicists have fully understood — it is easy to guide student - writers into becoming student-authors. We who teach Classics can add to our pedagogy the technological tools of the information economy, thus arming ourselves against charges of impracticality and at the same time possibly attracting students whose interests lie outside the Classics. And as digital libraries begin to inter-operate, they breathe new life into largely disregarded scholarly genres and invent entirely new ones — geographic information systems, computational linguistics, and so forth (Blackwell and Martin 2009).

Nonetheless, such calls it seems are only being *partially* heard, even in terms of much less "radical" innovation. Recent research by Dimitrios Vlachopoulos has investigated the perceptions of academic staff that teach classical languages (Greek and Latin) regarding the use of "online activities" in their teaching (Vlachopoulos 2009). In the first phase of this research, 33 instructors in Greece, Spain and the United States were first asked to

[535] The issue of multiformity, Homeric tradition and digital editions has been discussed earlier in this paper.

complete a three part survey that first asked them about their "digital profile" and their general level of information and communication technology (ICT) understanding. While the second part of the survey asked them to evaluate the potential of ICT in Classics and whether or not they or their students had the knowledge to actively utilize such technology, the third part asked instructors to outline the most significant challenges in using ICT for online course delivery. In the second phase of this research face-to-face interviews were conducted with about half of the participants, and Vlachopoulos emphasized that most of the participants were worried about the future of their departments and the amount of funding they received from their universities. "It was a common belief that new strategies need to be designed," Vlachopoulos reported, "in order to attract more students every year and to offer them more job opportunities."

The analysis of the survey and interview results led Vlachopoulos to classify the instructors into three groups: conservatives, who were completely closed to the use of innovative ICT in the classroom; mainstream, who even if they stated they were in favor of major changes in teaching, were "risk averters" and faced significant problems in deploying ICT; and early adopters, who were open to the innovative use of ICT in their classrooms. Despite the fears stated above about needing new methods to attract students, 46% of the group fell into the mainstream with 30% classified as conservatives and only 24% identified as early adopters. [536] Vlachopoulos stated that while early adopters wanted to create new roles in the classroom, explored new teaching methods with technology and reported a high level of willingness to pursue experimentation, mainstream faculty wanted "proven applications of recognized value" before they deployed them in their classroom and also needed significant technical support for almost all ICT application. Explaining this group classification further, Vlachopoulos detailed that:

> Only 15% of the instructors can be identified as early adopters concerning their skills in using ICT for learning activities. These individuals have studied computer science for personal use and use ICT every day in their personal life and almost every class they give. The majority of the instructors (55%) belong to the mainstream category since they haven't studied computer science and use ICT occasionally at home. In their classes they often use simple ICT applications, such as PowerPoint presentations, email and internet (Vlachopoulos 2009).

The largest area of support for the use of ICT was in terms of combining it with traditional teaching methods, since 70% of the instructors believed this was possible. In order to encourage the greater use of ICT within classical teaching, Vlachopoulos suggested that the designers of innovative projects would need to come up with strategies to attract more mainstream faculty but also cautioned that administrators would have to consider the greatly idiosyncratic nature of teaching within classics before deploying new teaching methods using ICT. In addition, as only 5 of the interviewees considered themselves as technologically self-sufficient, Vlachopoulos surmised that universities would need to provide a large amount of technical support in order to successfully deploy ICT in the classroom. As a final thought he noted that one of the most important points for encouraging more mainstream faculty to adopt innovative uses of ICT in teaching would be to convince them of its efficiency.

Another discovery highlighted by Vlachopoulos as part of his research was that he was not able to "find any department of Classics that applies a complete online language course in its curriculum." While some universities that were open to the use of information technology had designed online activities such as exercises, quizzes and surveys, there was "no complete course delivery with periodic and stable interaction between the members of a virtual community/classroom" (Vlachopoulos 2009).

An earlier JISC funded survey by the Higher Education Academy, History, Classics and Archaeology Subject Center (HCA) [537] pursued similar research and examined the use of e-resources in teaching and learning in the disciplines of history, classics and archaeology in the United Kingdom (MacMahon 2006). This survey made use of an online questionnaire, semi-structured interviews, and focus groups. The five most used e-resources were email, websites of their home institution, PowerPoint, e-journals, and other institutions' websites. Interestingly, the survey found that there was a significant difference between the e-resources that were the most

[536] This appears to confirm the earlier findings of the CSHE study (Harley et al. 2006b) that classicists use digital resources in the classroom less frequently than other disciplines.
[537] http://www.heacademy.ac.uk/hca

frequently used and those respondents reported they were most *likely* to use including software tools, e-books, digital archives, and virtual learning environments. One primary concern of faculty was the accessibility of the online learning materials. Another insight offered was that faculty often felt that an e-format was not always the best way of delivering what they considered to be essential learning materials for their teaching. Other areas of concern were digital rights issues, student competence to use electronic resources both in terms of IT skills and disciplinary knowledge, and low levels of institutional support for using such resources.

Nonetheless, the study authors reported that the responses to the questionnaire had convinced them that the use of e-resources had made a significant impact on teaching practices within the surveyed disciplines. The two alterations in teaching practice that were most frequently reported were an alteration in the learning materials and the teaching methods used to deliver them. Surveyed faculty also reported a number of positive and negative impacts of using e-resources on student learning. Access to a wider range of source materials was highly cited as a positive development, particularly since it enabled students to *conduct research* at an earlier stage in their education with both visual and textual materials, and faculty hoped that this would encourage independent learning. Faculty also noted that electronic resources permitted materials to be customized for the needs of different learning styles and accessed both off campus and all the time. On the other hand, some faculty feared that the rote use of e-resources would actually deter independent learning with a focus on "training" rather than education, that students would be discouraged from reading, and that students made over-use of the Internet and were also not discerning in their use of many questionable websites. The primary theme of these concerns was that e-resources should not replace face-to-face teaching. The creators of the survey thus concluded that "blended learning" or where e-resources formed part of their pedagogy best characterized the approach of faculty in classics, archaeology and history in terms of using e-resources.

At the same time, some other researchers have argued that not all applications of ICT within the classroom [538] necessarily need to be innovative or cutting edge to be *useful*. A recent article by Richard Ashdowne (2009) has examined the development of Galactica ("Greek and Latin Accidence Consolidation Training, Internet-Centered Assessment"), a tool that was designed to support the University of Oxford's Classics faculty language consolidation classes. Approximately 155 undergraduates begin classes at Oxford every year that require a knowledge of Greek, Latin or both, and the level of previous linguistic experience among students was found to vary greatly. Since intensive language classes were thus required for all these students along with frequent testing, Ashdowne observed that the department determined that some form of online evaluation in the key area of accidence testing would be highly desirable. "Moreover, most students now arrive with basic computing skills," Ashdown noted, "and it is the Faculty's stated view that it is an important part of degree-level education in Classics that students should develop relevant skills in using Classics-related electronic resources."

Consequently Galactica was developed to replace paper-based tests, and students were expected to log into this Internet-based system once a week for each language they were studying and to take the relevant tests. While Ashdowne stated that it was hoped that Galactica would provide classroom instructors with more time to focus on teaching and also assist students in developing the ability to manipulate polytonic Greek on a computer. Nonetheless, the online tests themselves, he noted largely shared the same purpose of the paper tests:

> …in sharing the aims of the paper-based system, Galactica illustrates how, although new technology *can* be used in new ways or for new ends, its application does not *have to* be pedagogically revolutionary. As they begin to develop, e-learning and e-assessment applications may often seem to focus on novelty (in the best sense) and innovation, creating educational tools to allow what would have been impossible or impractical before; but inasmuch as technology per se remains new in education, even traditional methods may be reinterpreted and implemented in a new way, as here. Classics is one field in which remembering the value of what has gone before is part of its intellectual core, and where rediscovering that value may itself be novel (Ashdowne 2009).

Ashdowne thus illustrated the important role that technology can play not only in helping classicists develop radical new teaching methodologies, but simply in helping them perform traditional teaching tasks such as evaluation in a far more efficient way. The Galactica system was based on TOIA (Technologies for Online

[538] While a full exploration of the use of ICT within classical teaching is beyond the scope of this review, for the development of one individual application for Latin please see (Mallon 2006) and for a general overview see (McManus and Rubino 2003)

Interoperable Assessment) [539] and required full Unicode compatibility, the ability to ask "grouped multiple-choice questions" as well as classroom management and result reporting. A variety of technical issues were encountered including the fact that TOIA was only compatible with of Internet Explorer for PCs. Another challenge was the lack of any recognized framework for "evaluating the pedagogical success of a system of this kind" (Ashdowne 2009). Nonetheless both student and instructor feedback on the system from the limited trails had been very positive. Ashdowne also declared that the minimal financial cost involved in developing Galactica also illustrated that "new technology can be used cost-effectively for very traditional purposes as well as for radically new ones." The main benefit of Galactica he ultimately concluded would be if it helped to free up class time in a cost-effective way. The efficiency of technology was thus recognized by both Ashdowne and Vlachopoulos as an important means of convincing traditional scholars to adopt a new tool.

One ambitious effort in the United Kingdom to encourage classicists not just to utilize digital resources within the classroom but also to actively participate in their design has been described by OKell et al. (2010). Between 2006 and 2008, the HCA [540] and the Centre for Excellence in Teaching and Learning for Reusable Learning Objects (RLO-CETL) [541] collaborated together to create a reusable learning object. Their project "digitally modeled the seminar (as a typical instance of humanities pedagogy) in a generic form inside a software package" and created the Generative Learning Object (GLO) Maker software [542] that could be used by faculty in their teaching. As OKell et al. explained, the RLO-CETL participated in this process for they wanted to "elicit pedagogical patterns" from various disciplines and then digitally model these patterns in ways that could be utilized by teaching practitioners. The RLO-CETL particularly wanted to ensure that the design process was "practitioner led" and that their domain expertise was recognized. The HCA participated in this collaboration out of a desire to engage with the e-learning community and to create more e-resources that would be appropriate for their disciplinary community. This collaboration illustrates the importance of domain specialists and technologists working together as well as the need to recognize domain expertise in the design of disciplinarily appropriate learning objects.

One key issue that the project wished to address was the need for students to engage in more critical learning. They cited surveys where university students expressed frustration at not being taught how to read texts and at lectures not giving them the "right answer." The project thus decided to focus on creating a learning object that supported students in learning to look at evidence, varying interpretations of that evidence, and to then make a critical argument of their own. The HCA brought a number of insights to this work from a JISC funded scoping survey they had conducted (MacMahon 2006) to determine the use of e-resources in teaching and learning in the UK in history, classics and archaeology. This survey illustrated that those faculty who participated supported "the creation of a community model" both to share their content [543] and to structure the pedagogy of the e-learning materials they used. Participants thought that their teaching would benefit from sharing e-learning resources with colleagues and they wanted customizable e-resources for particular content and learning objectives. At the same time, they did not want to require outside help or have to acquire new skills in order to be able to use e-learning resources. Similar results in terms of a desire not to need to learn any new skills to use digital resources was also reported by (Brown and Greengrass 2010, Warwick et al. 2008a).

A key research question of the project was to explore if the learning technology approaches that were used for scientific disciplines could also be used in the humanities. The HCA held a workshop with a number of academics where they reached the conclusion that the best approach would be to create a learning object that focused on an artifact and integrated interpretations of that artifact from different disciplines (a classicist, an

[539] http://www.toia.ac.uk/
[540] http://www.heacademy.ac.uk/hca
[541] http://www.rlo-cetl.ac.uk/
[542] http://www.glomaker.org/
[543] Interestingly, even though respondents overwhelmingly supported the *sharing* of e-resources, only 42% indicated that they were actually sharing e-resources with colleagues either within or outside of their home institution. The major reasons for this were a general lack of knowledge as to what types of e-resources were being used by their colleagues, a belief that learning materials should be "closely tailored" to particular learning objectives or course content, worries about ownership of materials, and lack of incentives to share. Personal contacts by far led to the most sharing of resources. Although there was some support for the creation of a repository or website to collect and make such e-resources searchable, there were great concerns about the sustainability of such a repository.

archaeologist and a historian). "The workshop participants had identified what humanities disciplines aim to do and the means by which they do it," OKell et al. explained, "This was achieved in a context where educational technologists keen to create the next generation of e-Learning resources could identify this aim and determine whether it could be modelled electronically" (OKell et al. 2010, pg. 158).

Thus this project actively sought to address the challenges of digitally modeling the pedagogical approaches of a particular discipline, by having disciplinary practitioners define a set of tasks and then having educational technologists see if they could successfully model them. In this case, the "powerful pedagogical pattern" that they modeled as a Generative Learning Object was that of "evaluating Multiple Interpretations" (*eMI*). JISC funded the development of a proof of concept software [544] and domain experts were involved for the entire process. After the altar of Pergamum was chosen as the artifact, a three-step process of storyboarding and refining ideas, mockup and digital design, and final implementation and testing was undertaken. The participating academics were asked to define questions that they wanted their students to be able to answer, and this resulted in three general types of questions: Origin, Purpose and Meaning.

Their attempt "to storyboard the learning process" faced a number of challenges for the scholars wanted to support both a linear (step by step from origin to meaning for each discipline) and branching navigation (e.g. comparing different disciplinary perspectives on the artifact's origin or meaning) through the module, but were uncertain if this was possible to design. While the original storyboard presented by scholars involved having students move sequentially through one discipline at a time in order to avoid confusion, the learning technologists suggested an alternative where students could compare multiple interpretations of each micro-theme such as "origin" in order to enable the comparison of multiple interpretations. This design choice was enthusiastically agreed upon and was consequently labeled "Access Views." In addition, as knowledge acquisition was a major goal of *eMI*, the module included various forms of multiple-choice questions to assess student learning.

A number of disciplinary audiences positively recognized the eMI module and OKell et al. concluded that by computationally modeling a specific pedagogical process the *eMI* framework could be easily *repurposed* by other groups designing digital learning objects. They also recognized, however, that there are limits to designing for reusability. "Some parts of the process can be noted and replicated to ensure useful outcomes," OKell et al. acknowledged, "but, overall, success when designing for reuse is dependent on the working relationship between the disciplinary practitioners driving the process and the learning technologists supporting them" (OKell, et al. pg. 167). The *eMI* project thus illustrated the importance of a good working relationship between information technologists and domain specialists for the long-term reusability of a digital object.

Looking Backward: State of Digital Classics in 2005

In 2005, the now defunct Arts & Humanities Data Service (AHDS) conducted a subject extension feasibility study to survey recent and current digital resource creation in areas not served by the AHDS including classics, philosophy and theology to see what level of service these disciplines might require from the AHDS. The report noted that both classics and ancient history were "relatively digitally mature and in need of advanced services." The report's author Reto Speck conducted a number of interviews with subject specialists in the field and also surveyed a number of digital projects. Speck noted that the digital projects in classics were exceptionally diverse as were the type of resource being digitized, including catalogues and bibliographies, prosopographical databases, manuscript images, papyri, inscriptions, artifacts, textual resources, line drawings, CAD and VR models of architectural structures, and spatial datasets. This wide variety of resources Speck noted helped to reflect the multi-disciplinary nature of classics, and also pointed out that many scholars who were interviewed, "suggested that ICT in general, and hypertext and hypermedia technology in particular, are beneficial to CAH research, since it enables the integration of textual, archaeology and historical sources and approaches into one research project."

[544] http://www.heacademy.ac.uk/hca/themes/e-learning/emi_glo

This ability of the digital medium to reintegrate the textual and material record and present a more sophisticated approach to exploring the ancient world was valued by many digital classics projects. Speck also found that the sophistication of computational methods used in projects varied greatly:

> For a large proportion of projects the digital component is clearly subsidiary to the wider research question and the computational methods employed are straight-forward; however, a significant minority of projects employs and devises advanced computational methods including multi-spectral imaging techniques, advanced 3-d modelling methods, and the development of generic and re-usable mark up schemes (Speck 2005).

A similar level of varying computational complexity was found in this project's survey of digital resources in classics. While some projects focused on using digital tools to better explore traditional questions, others were developing state-of-the-art tools to explore new questions.

Interestingly, Speck articulated that attempting to develop a single subject center to meet the needs of classics and Ancient History would likely fail to address both the inter-disciplinary nature of the field or the fact that most of the services requested of the AHDS were both quite specific and advanced. Nonetheless the report did offer six recommendations for supporting the needs of digital classics research: 1) "the development and promotion of generic methods and standards" such as TEI and EpiDoc; 2) the "integration and linkage of existing resources and cross-searching"; 3) the development of virtual research environments (VREs); 4) the "sharing of expertise, outcomes and methodologies and linking of projects"; 5) the need for national and international funding and 6) "information on encoding and display of non-Latin script." Of all these recommendations, many scholars stated that the main challenge for the future would be "in linking disparate collections of different data types to enable powerful cross-searching." In fact, a variety of projects have evolved to address just these issues such as Concordia, LaQuAT, and Interedition, all of which will be discussed in greater detail in the next section.

Looking Forward: Classics Cyberinfrastructure, Themes and Requirements in 2010

While the AHDS study of 2005 took a fairly broad approach to defining the needs of digital classics projects, three recent articles in a special issue of the *Digital Humanities Quarterly (DHQ)* that was dedicated to the theme: "Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure," have taken an even more expansive approach to the question of developing a cyberinfrastructure for digital classics, classics and the humanities as a whole. While Crane, Seales and Terras (2009) looked at the cyberinfrastructure requirements for classical philology as a means of exploring larger issues of digital classics, Crane et al. (2009a) summarized the challenges facing classical studies in the million book libraries being created by mass digitization projects, and Blackwell and Crane 2009 offered a conclusion to this special issue and an overview of its larger themes. Each of these articles and the requirements they list for a cyberinfrastructure for classics will be considered here.

While the theme of the advanced nature of computing in classics has been documented throughout this research, Crane, Seales and Terras (2009) suggest that this very level of "advancement" may present unexpected consequences:

> The early use of digital tools in classics may, paradoxically, work against the creative exploration of the digital world now taking shape. Classicists grew accustomed to treating their digital tools as adjuncts to an established print world. Publication — the core practice by which classicists establish their careers and their reputations — remains fundamentally conservative (Crane, Seales and Terras 2009).

They consequently recommended that philologists and indeed all classicists should move away from creating specialized *software* and start creating specialized *knowledge sources*; they envision a new digital infrastructure that supports the rethinking of all the traditional reference sources of classical studies.[545] The greatest barriers to be faced in creating this new infrastructure are social rather than technical as indicated by the fact that no

[545] The importance of digital reference works in an integrated research environment has also been recognized by de la Flor et al. (2010) in their discussion of developing the VRE-SDM: "Moreover, classicists frequently reference other material such as prior translations, dictionaries of Roman names and historical documents, whilst examining a manuscript. It would therefore be useful to be able to juxtapose the texts and notes they are working on with other paper and electronic materials, including being able to view partial transcriptions of the text alongside an image"

traditional elements of the scholarly infrastructure including commentaries, editions, grammars and lexica have truly been adapted to the digital world by being made machine actionable. Other problems include the fact that most scholarship is still single authored, the TLG provides digital texts without any critical commentary, and most major new critical editions have copyrights that remain with their publisher, thus leading to an over-reliance on the TLG.

Nonetheless, Crane, Seales and Terras advise that a cyberinfrastructure for philology and classics is slowly emerging and builds upon three earlier "stages of digital classics: incunabular projects, which retain the assumptions of print culture, knowledge bases produced by small, centralized projects, and digital communities, which allow many contributors to collaborate with minimal technical expertise." For digital incunabula, the TLG and the *Bryn Mawr Classical Review* are listed, the Perseus Digital Library is suggested as a knowledge base, and the Stoa Consortium is a model digital community. More importantly, the authors contend that these three classes of projects also reflect three separate sources of energy: "industrialized processes of mass digitization and of general algorithms, the specialized production of domain specific, machine actionable knowledge, and the generalized ability for many different individuals to contribute." The authors posit that when these three sources interact with each other they provide a new digital environment that makes possible ePhilology, eClassics and Cyberinfrastructure. Yet at the same time, they note unfortunately that our current infrastructure is not yet at this stage:

> The infrastructure of 2008 forces researchers in classics and in the humanities to develop autonomous, largely isolated, resources. We cannot apply any analysis to data that is not accessible. We need, at the least, to be able gather the data that is available today and, second, to ensure that we can retrieve the same data in 2050 or 2110 that we retrieve in 2010…. We need digital libraries that may be physically distributed in different parts of the world but that act as a single unit…(Crane, Seales and Terras 2009).

This quote illustrates the continuing challenges of limited access to primary sources and secondary scholarship, sustainable digital preservation, and creating an integrated user searching experience across virtual collections of data. The importance of an integrated infrastructure for research by classicists has also been recognized by the VRE-SDM project:

> The aim of the VRE-SDM project has been to construct a pilot of an integrated environment in which data (documents), tools and scholarly *instrumenta* could be available to the scholar as a complete and coherent resource. Scholars who edit ancient documents are always dealing with damaged or degraded texts and ideally require access to the originals, or the best possible facsimiles of the originals, in order to decipher and verify readings, and also to a wide range of scholarly aids and reference works (dictionaries, name-lists, editions of comparable texts, and so on) which are essential for interpretation of their texts (Bowman et al. 2010, pg. 90)

As Bowman et al. explain, an integrated research environment or cyberinfrastructure will require not just access to primary sources, but to digital tools and to a wide range of pre-existing reference tools/works that will need to be adapted for the digital environment. They also noted that many of the necessary collections have already been created or digitized but are unfortunately scattered across the websites of various museums and libraries.

In order to a more integrated classical cyberinfrastructure, Crane, Seales and Terras propose a minimum list of necessities, including: libraries or repositories that can provide sustainable preservation, "sophisticated citation and reference linking services," new forms of electronic publication, new models of collaboration, and a digital infrastructure that is portable across languages (Greek, Latin, Chinese, Arabic, etc.). They then conclude with three strategies to begin building this infrastructure: 1) optimizing machine translation for the field of classics 2) converting as much information as possible into machine actionable data and 3) using canonical literary texts that have already been marked up to serve as databases of linguistic annotations.

Crane et al. (2009a) provides an overview of the opportunities and challenges faced in moving from "small, carefully edited and curated digital collections to very large, industrially produced collections" with a focus on the role of classical collections and knowledge sources. The authors stress the need to create a classical *apographeme* online as an analogy to the genome, or the need to represent online:

> …the complete record of all Greek and Latin textual knowledge preserved from antiquity, ultimately including every inscription, papyrus, graffito, manuscript, printed edition and any writing bearing medium. This *apographeme* constitutes a superset of the capabilities and data that we inherit from print culture but it is a qualitatively different intellectual space (Crane et al. 2009a).

This argument focuses on the need to represent all Greek and Latin sources online in an integrated environment, whether inscribed on stone or printed in a book. Matching these new online collections with advanced OCR and other applications, Crane et al. (2009a) explain, is currently supporting a number of important new services, including the creation of multitexts, chronologically deeper corpora, and new "textual forms of bibliographic research." In this new world, the authors argue, all classicists are also acting as corpus linguists.

A large part of this paper is dedicated to outlining the services required for humanities users in massive digital collections, including access to physical images of sources, transcriptional data, basic page layout information, semantic markup within a text, dynamically generated knowledge, and finally, "linguistically labeled, machine actionable knowledge." The importance of access to "machine actionable knowledge" and the need for creators of digital classics resources to create data and sources that help build this knowledge base is a preeminent theme of this paper. But this process is two-fold as Crane et al. (2009a) explicate, for while scholars need to *create* data that can be used by *automatic* processes, they also need to be able to build *off* of data created by these processes as well.

The authors thus call for the creation of "fourth-generation collections" that will support a cyberinfrastructure in classics. Such collections will have a number of features: 1) they will include images of all source writing including papyri, inscriptions, manuscripts, and printed editions; 2) they will "manage the legacy structure of books;" 3) they will integrate XML transcriptions as they become available with image data, so that "all digital editions are, at the least, reborn digital"; 4) they will contain "machine actionable reference works" that are embedded in growing digital collections that automatically update themselves; 5) they will learn from their own data and collections; 6) they will learn from their users, or rather, contain automated systems that can learn from the annotations of their users; 7) they will adapt themselves to their readers either through watching their actions (personalization) or through user choice (customization); and 8) they will support "deep computation" with as many services as possible that can be applied to their content. As one of their final thoughts, the authors reiterate the point that a cyberinfrastructure for classics should include images of writing from all types of sources. "In a library grounded on images of writing," Crane et al. (2009a) suggest, "there is no fundamental reason not to integrate, at the base level, images of writing from all surfaces."[546] In fact, the difficulties of this integration of writing from the printed and material records will likely be one of the greatest technical challenges in developing a cyberinfrastructure for classics.

The conclusion of the special *DHQ* issue by Blackwell and Crane (2009) offered a summary of the various issues raised throughout and returned to the concepts of ePhilology, eClassics, and cyberinfrastructure. Any cyberinfrastructure for classics they argued must include open access data, comprehensive collections, software, "curated knowledge sources" and "advanced, domain optimized services." The authors put forward that any cyberinfrastructure for the humanities can easily begin with classics because not only is it one of the most digitally mature fields but for a variety of other reasons as well. First, classical studies provides a cultural heritage that is truly international. Second, although most of the *DHQ* articles in this special issue focused on the textual record, there is a vast body of untapped data about the ancient world in *archaeology*:

> The study of the Greco-Roman world demands new international practices with which to produce and share information. The next great advances in our understanding of the ancient world will come from mining and visualizing the full record, textual as well as material, that survives from or talks about every corner of the ancient world (Blackwell and Crane 2009).

Such a record can only be built through international collaboration. Third, the textual corpus of classics may be finite, but it has had an immense impact on human life. Fourth, "Greco-Roman antiquity demands a general architecture for many historical languages" so that technical development in supporting these languages can help lead to advances in supporting languages such as Sumerian and Coptic. Fifth, most contemporary scholarship is multi-lingual, and classics is truly one of the most fundamentally multilingual communities in the academy.[547] Sixth, knowledge and understanding of the extent of the Greco-Roman world could help lead to new

---

[546] This argument was also seen throughout this review, see in particular (Roueché 2009) and (Bagnall 2010)

[547] The challenges of developing a digital collection infrastructure that can accommodate a multilingual collection (Latin, Greek, Arabic and Italian) of both classical and medieval texts in the history of science, has been examined by the Archimedes Digital Library (http://archimedes.fas.harvard.edu/), and also see (Schoepflin 2003)

involvement with areas such as the Middle East in terms of this shared heritage. Seventh, "classical scholarship begins the continuous tradition of European literature and continues through the present." This is important, the authors note, for:

> An infrastructure that provides advanced services for primary and secondary sources on classical Greek and Latin includes inscriptions, papyri, medieval manuscripts, early modern printed books, and mature editions and reference works of the 19th and twentieth centuries. Even if we restrict ourselves to textual sources, those textual sources provide heterogeneous data about the ancient world. If we include the material record, then we need to manage videos and sound about the ancient world as well (Blackwell and Crane 2009).

Classics is such a broad discipline that the various infrastructure challenges it raises will also be important for the development of any larger cyberinfrastructure for the humanities. The final reason Blackwell and Crane give for letting classics help define the development of a broader cyberinfrastructure is that classicists have devoted at least a generation to developing tools and services and now "need a more robust environment and are ready to convert project-based efforts into a shared, permanent infrastructure" (Blackwell and Crane 2009)

In order to move from project-based efforts to a shared digital infrastructure, the authors list numerous specialized services developed by individual digital classics projects that will need to be supported, including: *canonical text services, OCR and page layout, morphological analysis, syntactic analysis, word sense discovery, named entity analysis, metrical analysis, translation support, CLIR, citation identification, quotation identification, translation identification, text alignment, version analysis and markup projection.* In addition to these services, two specific types of texts are required to support ePhilology in particular: 1) *Multitexts*: or "methods to track multiple versions of a text across time"—these methodologies allow for the creation of "true digital editions" that include all images of their source materials, various versioned and reconstructed editions, and multiple apparatus critici that are machine actionable. 2) *Parallel texts*—extends the idea of a multitext across languages and parallel texts. Other collections required to support ePhilology include wordnets, treebanks, linguistic annotations, machine actionable indices and commentaries.

Blackwell and Crane (2009) end their piece with thoughts on what is needed for true digital publication in cyberinfrastructure and the announcement of the Scaife Digital Library (SDL). The authors convincingly assert that: "just because information is on-line does not mean that that information has exploited the full potential of the digital medium" (Blackwell and Crane 2009). Classical materials they argue need to be available in interoperable formats and with open licenses (e.g. almost all of the TEI-XML texts in the Perseus Digital Library have been available for downloaded under a CC license since 2006.) Similarly the Center for Hellenic Studies (CHS) announced a plan in 2008 to create a digital library of new TEI compliant XML editions "for the first thousand years of Greek."

In order for an item placed online to be *useful* in a digital world, Blackwell and Crane propose that it must meet four conditions of digital scholarly publication: 1) the content must be of interest to those other than its creators 2) it must have a format that can be preserved and used for a long period of time 3) it needs at least one long-term home 4) it must be able to circulate freely. All objects that will be placed in the SDL must meet these requirements, and the authors also state that the SDL will not provide *services* to its end users, but rather provide access to re-purposable digital objects. In their final conclusion, Blackwell and Crane outline three issues to be faced or perhaps accepted, first, that in this new world, "all classicists are digital classicists" or at least they must become so in order for their scholarship to retain meaning; second, that classicists will need to work with scholars who have more advanced understanding of technology; and third, that new institutions are necessary, or a new hybrid library-publisher that can help classicists create and maintain their objects/services.

These articles illustrate a number of important issues to be considered for a classics cyberinfrastructure or indeed for a digital repository or federated series of repositories to meet the needs of digital classicists. These requirements include: open data and collections (open not just in terms of access, but in terms of openly licensed where all the data is available), curated knowledge sources and machine actionable reference works, both general and domain specialized services, collaboration both within the discipline of classics and with other disciplines, and an infrastructure that will support both a reasonable level of domain customization while still being flexible enough to provide general storage and high speed access to computational processes. Similarly, Mahony and Bodard (2010) have offered a similar list of requirements including "Digital infrastructure, Open

Access publication, re-use of freely licensed data and, and Semantic Web technologies" in order for Classics to fully engage with an "increasingly digital academic environment" (Mahony and Bodard 2010, pg. 5). The next section will outline a number of projects that have taken some initial steps towards building parts of this infrastructure.

Classics Cyberinfrastructure Projects

While there are a large number of national and international cyberinfrastructure projects that will be discussed in the next section of this report, a number of smaller projects have focused on providing greater integration of major digital classics resources or greater infrastructure for classics, sub-disciplines of classics or medieval studies. Some of these projects have been discussed in brief above. Some projects have been completed while others are still ongoing.

APIS—Advanced Papyrological Information System

This project has been discussed in greater detail in the Papyrology section.

CLAROS—Classical Art Research Center Online Services

The CLAROS [548] project will support the "virtual integration of digital assets on classical art" including pottery, gems, sculpture, iconography, and antiquaria. CLAROS is using "Semantic Web data integration technologies and state-of-the art image recognition algorithms" and seeks to bring classical art "to anyone, anytime, anywhere." [549] Its partner institutions include the Beazley Archive at Oxford, the German Archaeological Institute (DAI) in Berlin, the Lexicon of Greek Personal Names (LGPN), the Lexicon Iconographicum Mythologiae Classicae (LIMC Basel and LIMC Paris) and the Research Archive for Ancient Sculpture Cologne (Arachne). Currently no searching or browsing features are available, but the integrated database is supposed to become available in 2010. Recently the project has created a wiki [550] that includes descriptions of the RDF/XML CIDOC-CRM format and CLAROS entity description templates for Objects, Places, Periods, and People. CLAROS has also recently announced the MILARQ project that will run during 2010 and seeks to "enhance the execution speed of queries against the CLAROS data web." The project's goal is by the end of October 2010 to have enhanced the performance of the CLAROS data web to the point that it is ready for public release. The project website notes that this will be accomplished by enhancing Jena "the widely used open source Semantic Web data management platform employed by the CLAROS data web, specifically the creation of multiple indexes over the underlying RDF triple store, Jena TDB, and other optimizations relating to filter performance, thereby speeding the execution of more complex SPARQL queries against the stored data."

Concordia

The Concordia initiative [551] has been established by the Center for Computing in the Humanities at King's College, London and the ISAW at New York University. It is a "a transatlantic collaboration" that will support "dissemination of key epigraphical, papyrological and geographic resources for Greek and Roman culture in North Africa, and piloting of reusable, standard techniques for web-based cyberinfrastructure." [552] A number of major projects are included in this effort including the Duke Data Bank of Documentary Papyri, Epigraphische Datenbank Heidelberg (EDH), Inscriptions of Aphrodisias (2007), Inscriptions of Roman Cyrenaica, Inscriptions of Roman Tripolitania and Pleiades. Designed as a demonstration project, Concordia will unite these separate digital collections of inscriptions and papyri (that include 50,000 papyrological and 3,000 epigraphic texts) with the geographic dataset of Pleiades. Some newly digitized content will also be included such as 950 epigraphic texts. Concordia will use basic web architecture and standard formats (XHTML, EpiDoc/TEI XML, and Atom+GeoRSS). The main goal is to provider users with one textual search across these

[548] http://www.clarosnet.org/index.htm, and for a discussion of their
[549] A discussion of CLAROS and its potential for expanding access to classical art see Kurtz (2009).
[550] http://www.clarosnet.org/wiki/index.php?title=Main_Page
[551] http://concordia.atlantides.org/
[552] http://www.atlantides.org/trac/concordia/wiki/ProjectOverview

collections as well as "dynamic mapping and geographical correlation for arbitrary collections of humanities content, hosted anywhere on the web."

This project is set to conclude in 2010 and has created a project wiki that tracks deliverables, workshop "results and other general information. [553] A number of software tools have already been created including: epidoc2atom (a set XSLT sheets for "creating web feeds from EpiDoc conformant XML documents"), the Concordia Matchtool, a "framework for defining and executing rulesets to effect matching of records in two datasets," and Concordia Harvester, "software for crawling and indexing Atom+GeoRSS feeds." Several important deliverables that the Concordia project also plans to create include Atom + GeoRSS web feeds for all papyri and inscription collections and the ConcordiaThesaurus, "a controlled vocabulary for expressing classes of relationships (or even assertions) between web-based resources in the context of Atom+GeoRSS feeds."

Digital Antiquity

This project has been described in greater detail in the Archaeology subsection.

Digital Classicist

This project has been discussed in greater detail in the section on Open Access.

eAQUA

eAQUA [554] is a major German project that seeks to use NLP techniques such as text mining to generate "structured knowledge" from ancient texts and to provide this knowledge to classicists through a portal. Researchers in classics and computer science are working together on six sub-projects (Büchler et al. 2008).

1) Atthidographers—This subproject will use text mining methods to search through digital Greek corpora to try and discover previously unfound citations to and quotations of this group of annalistic and fragmentary Greek historians.

2) Reception of Plato's texts in ancient world—A combination of visualization and text mining techniques will be used to discover and graph quotations and citations of Plato in ancient texts (Büchler and Geßner 2009).

3) The meter of Plautus—This subproject will use NLP techniques to perform metrical analysis on the texts of the Latin poet Plautus (Deufert et al. 2010).

4) Knowledge Map of the Early Modern Period—This subproject extends the work of MATEO, CAMENA and Termini, [555] a collection of Latin books and tools to analyze them from the early modern period, and will explore new research using co-occurrence analysis and text mining to track lexical changes over time from the ancient to modern world as well as to create semantic views of the corpora.

5) Epigraphical work—Extraction of templates for inscriptions.

6) Papyrology—This subproject will use text-mining techniques to provide text completion for distributed fragmentary collections.

The eAQUA project also sponsored a full day workshop at the Digital Humanities 2010 conference on text-mining in the humanities. [556]

eSAD—e-Science and Ancient Documents

eSAD [557] or "Image, Text, Interpretation: e-Science, Technology and Documents project" is using computing technologies to aid classicists and other scholars in the task of reading ancient documents. This four year

[553] http://www.atlantides.org/trac/concordia/wiki
[554] http://www.eaqua.net/en/index.php
[555] http://www.uni-mannheim.de/mateo/camenahtdocs/camena.html
[556] http://dh2010.cch.kcl.ac.uk/academic-programme/pre-conference-workshops/workshop-2.html
[557] http://esad.classics.ox.ac.uk/

project has been undertaken by the University of Oxford with input from University College London and runs until the end of 2011. eSAD has two major research projects: 1) creating tools to aid in the reading of damaged texts such as stilus tablets at Vindolanda and 2) discovering how an Interpretation Support System (ISS) "can be used in the day-to-day reading of ancient documents and keep track of how the documents are interpreted and read." This project has published extensively on their work including (de la Flor et al. 2010, Olsen et al. 2009, Roued 2009, Roued-Cunliffe 2010, Tarte et al. 2009, Tarte 2010) and further discussion of these articles can be found in the Papyrology section.

Integrating Digital Papyrology & Papyri.info

Integrating Digital Papyrology (IDP) [558] is a project that was first conceived in 2004/5 when the Duke Data Bank of Documentary Papyri (DDBDP) and the Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV) began "mapping their two largely overlapping data-sets--Greek texts and descriptive metadata, respectively--to each other." In 2007, the Mellon Foundation provided initial funding to migrate DDBDP from SGML to EpiDoc and from betacode to Unicode Greek, to merge mapped DDBDP texts and HGV metadata in a single XML stream, and to then map these texts to their APIS records, including metadata and images. They also wished to create an enhanced papyrological navigator (PN) to support searching of this newly merged and mapped dataset. In October 2008, a new two year project was funded by Mellon, IDP-2, to "1) improve operability of the PN search interface on the merged and mapped data from the DDBDP, HGV, and APIS, (2) facilitate third-party use of the data and tools, (3) and create a version controlled, transparent and fully audited, multi-author, web-based, real-time, tagless, editing environment, which — in tandem with a new editorial infrastructure — will allow the entire community of papyrologists to take control of the process of populating these communal assets with data." The ultimate goal of the IDP is to create an editorial infrastructure where papyrologists can make contributions to this integrated knowledge source. The project wiki also provides extensive software descriptions and downloadable code. [559]

The related Papyri.info [560] website provides two major features: a list of links to papyrological resources and "a customized search engine (called the Papyrological Navigator) capable of retrieving information from multiple related sites." The Papyrological Navigator currently retrieves and displays information from the APIS, DDBDP and HGV. The goal of this project is to demonstrate "that a system can be designed to provide an integrated display of a variety of scholarly data sources relevant to the study of ancient texts." This prototype uses portlet technology, a higher resolution image display platform, and "moves beyond the creation of centralized "union databases," such as APIS, to leverage and integrate content created and hosted elsewhere in the scholarly world." A major research effort of this project is investigating the scalability of their approach, and they hope to design a system that could include and integrate data sources beyond the initial ones in this project. A portlet platform was also chosen in order to support "personalization and profiling" so scholars can use it efficiently in their research. A sample record [561] demonstrates the potential of this research by including the metadata for an individual papyrus (P.Oxy 4 744) from the APIS and HGV, with the full DDBDP transcription (with downloadable EpiDoc XML), an English translation (when available), and an image that can be focused in on in detail. More extensive technical documentation can be found at the IDP website. [562]

Interedition: an "Interoperable Supranational Infrastructure for Digital Editions"

The Interedition Project [563] has the major goal of promoting "interoperability of the tools and methodology" used in the field of digital scholarly editing. As the project website notes, many scholars have already created "amazing computer tools" and the goal of Interedition is to facilitate contact between scholars and to encourage creators of such tools to make their functionality open and available to others.

[558] http://idp.atlantides.org/trac/idp/wiki/
[559] http://idp.atlantides.org/trac/idp/wiki/IDPSoftware
[560] http://www.papyri.info/
[561] http://www.papyri.info/navigator/full/apis_toronto_17
[562] http://idp.atlantides.org/trac/idp/wiki/PapyrologicalNavigator
[563] http://www.interedition.eu/

This project is funded as an EU Cost Action from 2008 to 2012 and it will hold a series of meetings between researchers in the fields of digital literary research and IT to "meet on the topic of a shared supranational networked infrastructure for digital scholarly editing and analysis." At the end of this project, a roadmap will be delivered for the implementation of such an infrastructure. [564] They will also release a number of "proof-of-concept web services to demonstrate the viability of the ideas and concepts put forward by Interedition as a networked research platform." The Interedition project wiki [565] provides details about past and previous workshops and includes a list of four workgroups that have been created to work on the European dimension, prototyping, strategic IT recommendations and a roadmap. A draft architecture has also been proposed [566] and there is also a separate software development site for this project. [567]

LaQuAT—Linking and Querying of Ancient Texts

The LaQuAT [568] project was a collaboration between the center for e-Research at King's College London and the EPCC at the University of Edinburgh. The project explored the use of the OGSA-DAI data management software that is used to support "the exposure of data resources, such as relational or XML databases, on to grids" in the fields of epigraphy and papyrology. A small case study of integrating three digital classics resources, the HGV (also participating in the IDP project), Project Volterra, and the Inscriptions of Aphrodisias, was conducted and a demonstrator that searched across the three databases was created. The demonstrator is currently maintained by King's College but the ultimate plan is to make the infrastructure developed for this project part of DARIAH. As the data formats for all three databases were different, this project illustrated both the limitations and potential of linking up diverse data sets in the humanities. "More generally," Jackson et al. (2009) stated, "it was realised that once one starts joining databases, the fuzzy, uncertain, interpretative and inconsistent nature of the data that is generated by and used in humanities research leads to issues about the meaning of what is facilitated by linking these databases"(Jackson et al. 2009). One important conclusion of the LaQuAT project was thus for the need for *virtual* data centers that can integrate several resources while also allowing individual resources to maintain their *unique* formats.

Building A Humanities Cyberinfrastructure

Defining Digital Humanities, Cyberinfrastructure and the Future

Building a cyberinfrastructure for the humanities involves thinking both about digital humanities research and its current state as a discipline. In a recent article, Christine Borgman has outlined many of the challenges currently faced by the digital humanities community as it attempts to possibly come together as a larger discipline and struggles to plan for a shared cyberinfrastructure:

> The digital humanities are at a critical moment in the transition from a specialty area to a full-fledged community with a common set of methods, sources of evidence, and infrastructure – all of which are necessary for achieving academic recognition….Digital collections are proliferating, but most remain difficult to use, and digital scholarship remains a backwater in most humanities departments with respect to hiring, promotion, and teaching practices. Only the scholars themselves are in a position to move the field forward. Experiences of the sciences in their initiatives for cyberinfrastructure and eScience offer valuable lessons. (Borgman 2009).

Borgman maintained that in order for the digital humanities to be successful, scholars would need to begin moving more actively to both build the necessary infrastructure and to promote their own interests. Her article sought to serve as a call to action for scholars in the digital humanities with a focus on six factors that will affect the future of digital scholarship in the humanities: publication practices, data, research methods, collaboration, incentives, and learning, each of these themes will be discussed later in this section.

---

[564] http://w3.cost.esf.org/index.php?id=233&action_number=IS0704
[565] http://www.interedition.eu/wiki/index.php/Main_Page
[566] http://www.interedition.eu/wiki/index.php/WG2:Architecture
[567] http://arts-itsee.bham.ac.uk/trac/interedition/
[568] http://www.kcl.ac.uk/iss/cerch/projects/completed/laquat.html

Borgman provided a useful definition of the digital humanities as "a new set of practices, using new sets of technologies, to address research problems of the discipline." The digital humanities as a new set of practices and technologies also requires a particular type of infrastructure, and the requirements were first laid out concretely in 2006 by a report commissioned by the American Council of Learned Societies (ACLS):

> Cyberinfrastructure is more than the tangible network and the means of storage in digitized form. It is not only the discipline-specific software application and the project-specific data collections: it is also the more intangible layer of expertise and best practices, standards and tools, collections and collaborative environments that can be broadly *shared* across communities of inquiry (ACLS 2006).

The ACLS report also offered a list of characteristics that would be required of a humanities infrastructure: it will operate as a public good, be both sustainable and interoperable, encourages collaborative work and support experimentation. The definition and list of characteristics outlined above offer a number of major points that this report will consider in greater detail, particularly that infrastructure includes not only data collections and software for individual disciplines but also best practices, standards, collections and collaborations that can be shared across disciplines. The importance of *sharing* as key to the definition of infrastructure has also been expressed by Geoffrey Rockwell, who posited: "Anything that is needed to connect more than one person, project, or entity is infrastructure. Anything used exclusively by a project is not" (Rockwell 2010). In his own discussion of cyberinfrastructure, Rockwell seconded the point of the ACLS report that it should be broadly useful to the public, but also concluded that cyberinfrastructure needs to be "well understood enough" so that it is broadly useful, be able to foster economic or research activity, be "funded by the public for the public," be *invisible* so that its use becomes reliable and expected, and be maintained by a long-term organization. At the same time, Rockwell also made a key distinction between *humanities research* and *research infrastructure* or cyberinfrastructure, and the importance of supporting both. "Research, by contrast is not expected to be useful, necessarily, and certainly isn't expected to be useful to a public," Rockwell concluded, "Research is about that which we don't understand, while infrastructure really shouldn't be experimental"(Rockwell 2010). Rockwell's larger point was that in defining their cyberinfrastructure, humanists should remember that a "turn to infrastructure" involves political and sociological decisions and a possible redefinition of what is considered as "legitimate" research.

Another major point made by the ACLS report was that "extensive and reusable digital collections" were at the core of any cyberinfrastructure and that scholars should be engaged in the development of these collections, for as the commission noted, almost all successful tool building is dependent on both the existence of and access to digital collections. The concept of services, content and tools as *infrastructure* was seen repeatedly throughout not just this report but also in the discussions by TextGrid and other humanities cyberinfrastructure research projects, and will be explored further in the next section.

Open Content, Services and Tools as Infrastructure

A joint report by JISC-NSF in 2007 emphasized the idea that both content and the tools necessary to exploit it are essential components of infrastructure. "In the cyber age, collections of digital content and the software to interpret them have become the foundation for discovery;" Arms and Larsen (2007) insisted, "they have entered the realm of infrastructure." In their overview of the role of virtual research environments in scholarly communication, Voss and Procter (2009) offered a similar conclusion. "The concept of "content as infrastructure" emphasises the increasing importance of collections of research data as a reusable infrastructure," Voss and Procter explained, "that builds on top of the physical research computing infrastructure and traditional infrastructures such as scientific instruments or libraries"(Voss and Procter 2009).

At the same time the authors of the JISC-NSF report stated that more cultural heritage and educational organizations needed to work together to produce and share their content, concluding that: "The arduous goal of open access in the humanities can only be achieved when public institutions no longer invest in endeavors with proprietary output." Since openly licensed or content that is freely available for reuse is such a fundamental part of infrastructure, the ACLS offered a similar warning, suggesting that more universities needed to work to create, digitize and preserve their own collections either locally or consortially, rather than renting access to

materials. The Association of Research Libraries (ARL) has also recently made a similar call for large-scale government funding to create a "universal, open library or digital data commons" (ARL 2009a).

The creators of TextGrid have reached similar conclusions regarding the primacy of both *content* and *services* in what they have labeled eHumanities ecosystems as an alternative to the term cyberinfrastructure. "However, at least for eHumanities ecosystems a model in which services reign supreme and content is exclusively seen as the matter on which the services operate is not satisfactory," Ludwig and Küster articulate, "For eHumanities ecosystems need models in which both content and services are first-class citizens" (Ludwig and Küster 2008).

In order to address the need to provide services on an infrastructural level for digital humanities research, the HiTHeR (e-Humanities High Throughput Computing) [569] project sought to embed a self-organizing text mining application/agent as a RESTful web service in an "e-Humanities ecosystem." Blanke, Hedges and Palmer (2009) provided an overview of this project that sought to explore what "digital services and value-creating activities" will particularly serve e-Humanities research. Although the particular tool described sought to create an "automatic chain of readings" for the *Nineteenth Century Serials Edition Project* (NCSE), [570] the larger question considered was how such an agent could be integrated into a large humanities cyberinfrastructure. A "resourceful web service" approach was used in order to avoid creating yet another isolated tool or web site solution. One of the greatest challenges for tool developers in the digital humanities the authors thus declared was determining how to create tools that were both appropriate for the traditional research scholars may wish to pursue while still allowing innovative work. [571]

The text mining service offered by the HiTHer project created a "semantic view" or an automatically generated browsing interface to the NCSE text collection:

> HiTHeR will offer an interface to primary resources by automatically generating a chain of related documents for reading. Users of HiTHeR are able to upload collections and retrieve lists of reference documents in their collections together with the N most similar documents to this reference document (Blanke, Hedges and Palmer 2009).

Since HiTHeR also aimed to provide a comprehensive research platform, they chose to offer several text mining algorithms for their users in terms of creating this "chain of readings." In addition, users could upload their own documents, not just use this tool with NCSE collections.

The HiTHer project quickly discovered, however, that standard computing environments did not provide the level of processing power necessary to run these algorithms. To resolve this problem, they built an infrastructure based on high-throughput computing (HTC) that uses many computational resources to accomplish a single computational task. They made use of the Condor toolkit that then let them rely on two types of computers at King's College London, underutilized desktop computers and dedicated servers. The authors thus assert that HiTHeR "illustrates how e-Humanities centres can be served by implementing their own local research infrastructure, which they can relatively easily build using existing resources like standard desktop networks" (Blanke, Hedges and Palmer 2009).

Another insight offered by the HiTHeR research group was that for most applications in the humanities, "large computing power will only be needed to prepare data sets for human analysis" (Blanke, Hedges and Palmer 2009). They suggested that for much humanities research, a user would simply need to call on heavy processing power to analyze a data set *once* and then would want to spend the rest of their time accessing and analyzing the results, or in other words most humanists would need a "create once-read many resources" application environment. This led them to ultimately deploy HiTHer as a restful web service where humanities scholars could call upon a variety of text mining algorithms and then receive the results in a variety of formats (XHTML, Atom, etc.)

---

[569] http://hither.cerch.kcl.ac.uk/

[570] http://www.ncse.ac.uk/

[571] In a discussion of his annotation tool Pliny, John Bradley made similar claims and reiterated a point he had also made earlier (Bradley 2005), "that tool builders in the digital humanities would have better success persuading their non-digital colleagues that computers could have a significant positive benefit on their research if the tools they built fit better into how humanities scholarship is generally done, rather than if they developed new tools that were premised upon a radically different way to do things" (Bradley 2008).

The importance of services, or digital tools, more specifically, as infrastructure has also been discussed by Geoffrey Rockwell who provided an overview of the development of infrastructure for textual analysis, which included the creation of a portal for textual research called TAPoR and the development of a set of reference tools TAPoRware. [572] The general model intended was that this portal could be used to discover and use tools that had been registered by their creators as web services that were running in various locations. The portal was intended both to provide scholars easy access to already existing tools and to support the registration, creation and publishing of new services. Currently the portal is being reinvented, Rockwell reported, since many scholars did not find it easy to use, and he also suggested that web services are often not as reliable as they should be and most users require both simplicity and reliability. "My point here is that the model was to keep tool development as research but make the research tools easy to discover and use through portal-like infrastructure," Rockwell explained, "A further paradigm was that tools could be embedded in online texts as small viral badges, thereby hiding the portal and foregrounding the visible text, an experiment we are just embarking on" (Rockwell 2010). While Rockwell accentuated that digital tools were an important part of the portal infrastructure that at times needed to be "invisible" in order to make the content primary, he also argued that tool development is an important part of the humanities research process in itself.

Research libraries and digital repositories, as potential key components of cyberinfrastructure for the humanities, will also need to address the complexities of providing access to both *content* and *services* as part of a larger networked infrastructure according to a recent ARL report on digital repository services for research libraries:

> …managing unique content, not just traditional special collections but entirely new kinds of works and locally-created content, will be an important emphasis for collection and management. As users exercise new capabilities and require new services, library services will become less "localized" within the library and within campus systems and expand into the general network environment. Library services increasingly mean machine-machine interactions and will be embeddable in a variety of non-library environments (ARL 2009b).

The services provided by digital repositories within research libraries will thus need to move beyond the individual library to encompass services required in a larger network environment and new content of all kinds will also be required to support the research needs of users. This report also made the important point that many "services" that will be required will be to support machine to machine communication through the use of relevant standards and protocols.

In addition to services and content, digital tools, as outlined by Rockwell above, are another key component of infrastructure. As explicated by Nguyen and Shilton (2008) in their survey of existing digital tools, digital tools are typically distinct from the other services and resources created by digital humanities centers. [573] They defined tools as "software developed for the creation, interpretation, or sharing and communication of digital humanities resources and collections." Nguyen and Shilton specifically evaluated the findability and usability of digital tools that were provided by digital humanities centers and created a typology that further defined tools according to their objectives ("access and exploration of resources," "insight and interpretation" or to find larger patterns and interpret them, to support creation of new digital resources, and "community and communication"), technological origins and associated resources. In this particular study they excluded tools developed outside the digital humanities community or that had been developed to function with only a single digital resource or collection. To further manage the scope of their research they also limited the concept of findability to the ability of a user to find a tool on a digital humanities center website.

At the same time, Nguyen and Shilton granted that a larger research study determining how easy it is for users to find digital tools using existing search engines and metadata structures would be very useful. In fact, the difficulty scholars have in finding relevant digital tools was recognized by the report of a 2009 workshop (Cohen et al. 2009) sponsored by the NSF, NEH and the Institute for Museum and Library Services (IMLS) that investigated what would be required to create an infrastructure for digital tools that could then support "data-driven scholarship." Nguyen and Shilton developed an evaluation framework to assess the strength of each tool

---

[572] http://portal.tapor.ca and http://taporware.cmaster.ca

[573] The research conducted by Lilly Nguyen and Katie Shilton, "Tools for Humanists" was part of a larger research study of digital humanities centers by Diane Zorich (Zorich 2008).

in terms of its easy identification, "feature, display, and access," the clarity of documentation or description, and ease of operation. The effectiveness or technical performance of tools was not evaluated. Of the 39 tools evaluated, only seven received high marks, and among the highest scoring tools were those such as Zotero [574] and Omeka, [575] both created by the Center for History and New Media [576] at George Mason University, and both of which have extensive documentation, technical support and devoted user communities. One feature of the highest-rated tools was their choice of words for "feature and display" that distinguished them as actual *tools*, and all tools fared better on variables that measured "ease of access" than "clarity of use." Nguyen and Shilton offered seven useful recommendations in terms of best practices for future digital tool designers: highlight tools more prominently on websites, offer a specific description of the tool's purpose and intended audience, make previews available (e.g. screenshots, tutorials, demos), provide technical support (FAQ, email address), clearly state the technical requirements for use or download, provide easy to use instructions on how to download a tool or interact with it (e.g. if tool is embedded in a Web browser), and perhaps most importantly, plan for the sustainability of a tool.

Further research by Shilton (Shilton 2009) explored the *sustainability* of digital tools in terms of the institutional support that existed to sustain the 39 tools identified in the previous study. Shilton proposed two new metrics "longevity of support" or the date a tool was established or other versioning information and "support for tool" defined as the level of technical support (e.g. number of updates, release timelines, open standards, long-term funding, etc.) provided for a tool. While acknowledging that infrastructure is a very broad term, Shilton explained that her report focused on the "aspects of infrastructure" that could be evaluated by examining tools public websites. Further research she argued should consider the more "subtle and intangible" aspects of infrastructure such as "human capital, dedication and institutional context." Another important dimension of tool value that she argued should also be explored in further research was the *utility* of a tool to humanities research.

Utilizing the above metrics, Shilton analyzed those of the 39 tools that still existed, and found that most of the tools that were rated highly in the first project also scored highly again in terms of sustainability. "The findings suggest that accessibility of tools and the quality of their supporting infrastructure," Shilton observed, "are, in fact, correlated. A successful combination of accessibility, longevity and support add to the *value* of a tool for researchers"(Shilton 2009). While some older tools had evolved into new ones, other tools had simply been "abandoned" due to loss of "interest, time or funding." Shilton thus offered a number of best practices in terms of sustainability including *website design* that makes tools easy to find and indicates that they are supported, and *professionalism*, or viewing tools not just as one time programming projects but as "products to support rigorous and long-term scholarship" that require both stewardship and ongoing institutional support. In agreement with Cohen et al. (2009), Shilton noted that developing a strong *user community* is a critical component of encouraging both tool accessibility and sustainability. Shilton also concluded, however, that a seamless and *invisible* cyberinfrastructure of digital tools for humanists was still only in its infancy. Nonetheless, she still proposed that "imagining the components of a curated infrastructure is an important next step for digital humanities research."

The workshop report by Cohen et al. (2009) listed a number of components that would be required for a *curated* infrastructure for digital tools and also delineated the problems with creating, promoting, and preserving digital tools such an infrastructure would need to address. To begin with, this report outlined the myriad problems faced by the creators of digital tools, including the *conceptualization* of the tool (e.g. what type of application should be built?), the ambiguous notions regarding what constitutes a tool, the failure of tool registries to gain builder participation, and the challenges of categorizing tools within taxonomies so that they can be found. Even after successful tool conceptualization, they stated, digital tool design projects still face issues with finding staff, community participation, and project management. In agreement with Shilton, Cohen et al. reported that

[574] http://www.zotero.org/
[575] http://omeka.org/
[576] http://chnm.gmu.edu/

much tool building did not meet acceptable levels of *professionalism*, with effective planning, version control of code, communication among staff and plans for long-term support.

Another important problem Cohen et al. described was how even after a tool has been successfully developed to the point where it can be distributed, there is the need to "attract, retain and serve users." The issues outlined above by Nguyen and Shilton regarding the findability, accessibility, and lack of transparency and documentation for tools, were also reiterated as barriers to building successful user communities. "In short, if concerns about the creation and production of tools has to do with the *supply* of new digital methods," Cohen et al. explained, "more has to be done on the other side of the equation: the *demand* for these digital methods and tools. User bases must be cultivated and are unlikely to appear naturally, and few projects do the necessary branding, marketing, and dissemination of their tool in the way that commercial software efforts do" (Cohen et al. 2009).

A variety of solutions were proposed to address these and other problems, but the discussion of most attendees according to Cohen et al. (2009) centered around cyberinfrastructure, albeit with various labels, including repository, registry, consortium and "invisible college." The major theme that emerged from these discussions was the need for an "economy of scale and the focusing of attention." A list of useful features for any infrastructure to be developed was also presented, including: a code depository, development management tools (team management, wikis, bug tracking), an outreach function to explain tools and methods, a discovery function, documentation support (including encouraging standardization), the running of contests or exchanges, *discipline-specific* code "cookbooks" and reviews, support to both develop and run training seminars, and resources to lobby for tool development and open access to content. Discussion of these features and a "draft strawman" RFP for a tool infrastructure that were circulated at this meeting raised a number of important issues Cohen et al. stated, the most important of which was *audience*, or who exactly would use this cyberinfrastructure, end-users or developers, or both? After considering the various models, the idea of the "invisible college" that focused on *communities* rather than "static resources" was agreed upon as a more viable solution than a repository. It was suggested that an invisible college approach might foster symposia, expert seminars and peer review of digital tools, as well as a system of incentives and rewards for "membership" in the college.

Behind all of these discussions, Cohen et al. pointed out was the recognition that all tool building and tool use must be deeply *embedded* within "scholarly communities of practice" and that such communities need to be promoted and be international in scope. The group ultimately envisioned a dynamic site similar to SourceForge that would provide 1) a "tool development environment"; 2) a "curated tools repository" that would provide peer review mechanisms as well as support discovery of tools; 3) functionality that supported both community building and marketing. Such a site they concluded might be linked up to the Bamboo Project, but they also acknowledged that their vision was a complex undertaking and thus proposed several themes for which to seek funding: the promotion of sustainable and interoperable tools, the creation of and support for infrastructure, and increasing rewards for the development of digital tools.

To promote sustainable and interoperable tools they proposed funding programs to train digital humanities trainers, to provide a grant opportunity for collaborative work that embedded already successful digital tools within a significant digital humanities collection, and to fund grant opportunities that would make two or more significant already existing tools interoperable. In order to promote the creation of infrastructure, they proposed securing grant funding to create a "shared tools development infrastructure" that should include developer tools, programming "cookbooks," and other relevant resources. Such an infrastructure they concluded, however, should not be "owned" by any individual or small group of universities, but might instead be hosted by centerNET or the Alliance of Digital Humanities Organizations (ADHO). Funding for such an infrastructure they insisted should also include a salary for an experienced "project management evangelist." In addition, they advocated funding a "Curated tools repository" or a form of digital tools review site or journal, as a means both of storing at least one copy of all tools submitted for publication and providing peer review mechanisms for evaluating those tools. Such a repository could provide discovery and recommender services, but would also

require a general editor and a strong board of respected digital humanists. Both of these suggestions illustrate the importance of staffing as a key component of any infrastructure.

New Evaluation and Incentive Models for Digital Scholarship & Publishing

While the ACLS report argued that young scholars would need to be offered more "formal venues and opportunities for training and encouragement" (ACLS 2006) in order to successfully pursue new digital scholarship, the CSHE report on scholarly communication practices found no evidence suggesting that technologically sophisticated graduate students, postdoctoral scholars or assistant professors were eagerly pursuing digital publishing opportunities in place of traditional venues. "In fact, as arguably the most vulnerable populations in the scholarly community," Harley et al. imparted, "one would expect them to hew to the norms of their chosen discipline, and they do" (Harley et al. 2010, pg. 9). Senior scholars who already had tenure were undertaking most of the greatest digital innovation they discovered. The irony remained, however, that young scholars were often expected by their senior colleagues to transform professions that would still as yet not recognize digital scholarship.

A recent article by Stephen Nichols has also criticized this lack of support for both *producing* and *evaluating* digital scholarship among the humanities:

> While attitudes more favourable to the needs of digital humanities projects are slowly evolving, we have yet to see a general acceptance of new approaches. Indeed, even where digital projects have been embraced, evidence suggests that attitudes from traditional or analogue scholarship continue to influence the way projects are evaluated, a practice that younger, untenured colleagues often find intimidating. At least as far as the demands of humanities credentialing are concerned, the dominion of the typewriter has yet to give way to that of the computer, metaphorically speaking (Nichols 2009).

This criticism was confirmed by the research of the LAIRAH project (Warwick et al. 2008b), which illustrated that many scholars creating digital humanities projects had received little support or recognition from their institutions. Further evidence of this trend was also demonstrated by a workshop at the 2010 Modern Language Association (MLA) Conference on assessing digital scholarship in non-traditional formats as detailed by Edmond and Schreibman (2010). When the first case study at the MLA workshop presented a digital edition of a little known poet complete with extensive scholarly apparatus, the first comment from a workshop participant was that the creation of such an edition was not *scholarship* but *service*. This attitude was reflected throughout the CSHE case study of archaeology as well, and many archaeologists that much digital work was a form of service but not scholarship. "The creation of a scholarly edition was a service activity, not a scholarly one regardless of the medium of presentation," Edmond and Schreibman reported, "The workshop facilitators immediately realized that the battle lines were far from fixed and that having a work in digital form only served to reinforce certain existing prejudices rather than allow for a widened scholarly horizon" (Edmond and Schreibman 2010).

One reason that Edmond and Schreibman (2010) suggested for the lack of "trust" in digital resources as real scholarship is that such resources are still perceived by many scholars as *ephemeral* and *transient*, and they thus proposed that the development of *sustainable* infrastructure for digital scholarship might help alleviate some of this distrust. In addition, Edmond and Schreibman also pointed out that even if the framework for print based scholarship was so "embedded in the academic and institutional systems that support it as to be nearly invisible" it was still an infrastructure in itself that many scholars had long since realized was beginning to break:

> One might have presumed that our non-digital colleagues might have looked to digital publication as a way out of the current difficulties; as a way of building new institutional structures to support in the first instance traditional research activities while exploring new models made possible by digital formats. But rather, the opposite has happened. There has arisen instead a bunker mentality clinging to the high old ways as assiduously as the British clung to the Ra (Edmond and Schreibman 2010).

Despite the many challenges facing traditional scholarly publishing, Edmond and Schreibman sadly acknowledged that many scholars still did not see any solutions to the "print crisis" through the digital publication of scholarship.

Borgman also affirmed in her article that neither journal nor book publishing in the humanities have rapidly embraced the digital world for a variety of reasons, including a distrust of online dissemination and an

unwillingness to try out new technologies. This needs to change, Borgman argued, because the "love affair with print" endangers both *traditional* and digital humanities scholarship. As print only publication continues to decrease, those who rely on it as the sole outlet for their scholarship Borgman concluded will be talking to an ever smaller audience. She also proposed that digital publishing offered a number of advantages over print, including the ability to incorporate dynamic multi-media or hypermedia, the possibility of reaching larger audiences, a far shorter time to publication, possibly heightened levels of citation, and easier access to digital materials. [577]

In addition, Borgman also stated that one key benefit of digital publishing for the humanities is that it "offers different ways of expressing ideas and presenting evidence for those ideas" (Borgman 2009). The ability of digital scholarship to be linked to not just the primary source data on which it is based but to be able to demonstrate different levels of scholarly *certainty* or highlight the *interpretative* nature of humanities scholarship was a factor that many digital classicists lauded as well [578] and was considered to be an essential component of any humanities infrastructure. Indeed, a number of archaeologists interviewed in the CSHE report argued that the major reason they would not consider websites as scholarly productions for tenure reviews was that few if any websites made a formal argument or offered an interpretative analysis of the evidence they provided (Harley et. al. 2010). Nonetheless developing infrastructure that not just supports but reflects the interpretative nature of the data it contains is a critical challenge. "The nexus between data gathering (or digitization) and interpretation," Stuart Dunn has argued, "is the crucial issue that librarians and technical developers are faced with when planning, or otherwise engaging with, the deployment of a VRE in archaeology, or indeed in the humanities more generally"(Dunn 2009).

A related issue addressed by Borgman is how to resolve several major *disincentives* she had identified as likely to prevent *traditional* humanities scholars from embracing open data and digital scholarship (Borgman 2009). Borgman stated that many humanists have various reasons for not wishing to either share their data or the products of their research. These reasons included the fact that there is often far more reward for *publishing* papers than *releasing* data, that the efforts to *documen*t one's data and sources for others is far more challenging than just for oneself, that not sharing data and sources can at times offer a *competitive* advantage to establish a priority of claims, and that many scholars view data as their own *intellectual property*. The CSHE report described a similar "culture of ownership" among archaeologists who were often reluctant to share data for fear of being "scooped." Borgman argues, however, that, each of these disincentives against sharing have some potential solutions. The reward structure for publishing rather than sharing is the most universal disincentive Borgman grant but she also argued that this environment is beginning to shift. In terms of data documentation challenges, Borgman proposed new partnerships between humanities scholars and information professionals. For the other disincentives, Borgman recommended short and sometimes long-term *embargoes* of data and publications that could serve to protect scholars rights for a time while also ensuring others will have access to the data eventually. Borgman acknowledged, however, that many scholars would also like to prevent access to their *sources* of data until after they have published. Nonetheless, she concludes that:

> As data sources such as manuscripts and out-of-print books are digitized and made publicly available, individual scholars will be less able to hoard their sources. This effect of digitization on humanities scholarship has been little explored, but could be profound. Open access to sources promotes participation and collaboration, while the privacy rules of libraries and archives ensure that the identity of individuals using specific sources is not revealed (Borgman 2009).

Borgman ultimately proposed that any infrastructure developed for the humanities should "err toward openness" in order to advance the field more quickly.

While the future of digital scholarship even within the smaller realm of digital classics is beyond the scope of this individual report, the challenges of gaining acceptance for digital classics projects and demonstrating how they in many ways can *enhance* traditional scholarship as well as support *new* scholarship were well illustrated

---

[577] Gabriel Bodard also offered a similar list of the advantages of digital publishing for classical scholarship (Bodard 2008) that was discussed earlier in this paper.

[578] Such as with Bodard and Garces (2009) in terms of digital editions, with Barker et al. (2010) for visualizations of historical narratives, with Beacham and Denard (2003), Flaten (2009), and Koll et al. (2009) for 3-D archaeological models and reconstruction

by the overview of digital classics projects earlier in this report. Despite the scholarly distrust of many digital publications highlighted by Edmond and Schreibman (2010) and Borgman (2009), peer review and the vetting of data were important components of many digital projects such as SAVE, Suda Online, Pleiades and Integrating Digital Papyrology.

Challenges of Humanities Data & Digital Infrastructure

"Central to the notion of cyberinfrastructure and eScience is that "data" have become essential scholarly objects," Borgman observed, "to be captured, mined, used, and reused" (Borgman 2009). Various types of data exist in both humanities and scientific research and Borgman listed several kinds of data including observational (surveys), computational data (from models or simulations), experimental data (laboratory work), and records (government, business, archival). While it is this last form of data that is used most frequently by humanists Borgman suggested, a fuller understanding of the nature of humanities data is a significant research challenge facing the designers of any cyberinfrastructure. Despite the belief that data in the sciences is very easy to define, Borgman also cited research from her own experience in environmental science where there were many "differing views of data on concepts as basic as temperature"(Borgman 2009).

Borgman criticized the fact that there were no significant "social studies of humanities" that would help better define the *nature* of data in humanities research:

> Lacking an external perspective, humanities scholars need to be particularly attentive to unstated assumptions about their data, sources of evidence, and epistemology. We are only beginning to understand what constitute data in the humanities, let alone how data differ from scholar to scholar and from author to reader. As Allen Renear remarked, "in the humanities, one person's data is another's theory" (Borgman 2009).

This lack of deeper understanding about the nature of humanities data raises complicated questions regarding what type of data is produced, how it should be captured and how it should be curated for reuse. Borgman also drew attention to Clifford Lynch's dichotomy of data as *raw material* vs. *interpretation* (Lynch 2002), pointing out that it brings up two relevant issues for the digital humanities. First, raw material is far more likely to be curated then scholars' interpretations of materials, and while it may be the nature of humanities research to constantly reinterpret sources, "what is new is the necessity of making explicit decisions about what survives for migration to new systems and formats." Secondly, humanities scholars usually have little control over the intellectual property rights of the sources they use (e.g. images of manuscripts, cuneiform tablets), a factor that can make data sharing very complicated in the humanities.

Another interesting comparison between the data practices of those working in the digital humanities and those working in the sciences was offered by Edmond and Schreibman (2010).

> If we apply a science paradigm, a digital humanities scholar could be compared to an experimental physicist, as someone who designs processes and instruments to find the answers to their research questions. But the most striking difference between the experimental humanist and the experimental physicist lies in the fate of these processes and instruments after the article on the findings they enabled has been written: they are transcended, perhaps licensed to another for further use, perhaps simply discarded. Why are we so different about our electronic data? Would it be enough for humanistic scholars as well to draw their conclusions and let it go either to be developed by someone else or to mildew? Or is there something inherently different in the nature of our data, that we should be so attached to its survival? For example, we expect to receive credit for scholarly editions—why should we not receive it for digital scholarly editions? Are the data collections created by humanists inherently more accessible and open than an experimental physicist's algorithm or shade-tree spectroscope? (Edmond and Schreibman 2010)

In addition to the unique nature of humanities data, Edmond and Schreibman also agreed with Borgman that there were many challenges to data reuse even beyond the frequently cited problem of intellectual property rights. They stated that once a scholar's colleagues and friends had typically looked at a digital project, the actual exploration and reuse of digital project materials was very low. Edmond and Schreibman speculated that the organization of a digital dataset or collection might appear to be too "powerful an act of editorialism" to many scholars for them to believe more original investigation can be conducted with the same materials. They also suggested, however, that the lack of *infrastructure* to communicate about digital works may also be greatly hindering their reuse. "Stripped of publishers' lists, of their marketing channels and peer review and quality

control systems," they wondered, "are we failing the next generation of scholars by creating too many resources in the wild?" (Edmond and Schreibman 2010).

The lack of formal dissemination and communication channels to promote digital resources is particularly problematic due to the sheer amount of potentially relevant data and tools (as well as irrelevant) that are available online. While the challenges of this data deluge are often discussed, particularly in terms of e-science, Stuart Dunn has suggested that for the digital humanities a more apt term might be the "complexity deluge":

> Driven by increased availability of relatively cheap digitization technologies and the development of software tools that support both existing research tasks and wholly new ones, the digital arts and humanities are now facing what might be termed a complexity deluge. This can be defined as the presence of a range of opportunities arising from the rate of technological change and the availability of e-infrastructure, that the mainstream academic community is not yet equipped to address its research questions with (Dunn 2009).

Dunn worried that this lack of readiness on the part of academia meant that *technology* rather than "research questions" would drive the development of infrastructure agendas

The complex nature of humanities data and the challenges of building a cyberinfrastructure for it have also been explored by Blanke et al. (2008, 2009), particularly in terms of the *semantically* and *structurally* diverse data sets that are involved and the highly *contextual* and *qualitative* nature of much of the data. "The integration of data items into arts and humanities research is non-trivial, as complicated semantics underlie the archives of human reports," they explained, "Humanities data may be highly contextual, its interpretation depending on relationships to other resources and collections, which are not necessarily digital"(Blanke et al. 2009). This semantic, structural, and contextual complexity led to a number of computational problems including the lack of formats or interfaces to make data/systems interoperable (Blanke et al. 2008). The difficulty of designing for the "fuzzy" and inconsistent nature of data in the humanities was also acknowledged by OKell et al. (2010) in their overview of creating a reusable digital learning object and they consequently labeled the humanities as an "ill structured knowledge domain."

Similar challenges in humanities data integration were also reported by the LaQuAT project when they evaluated the results of integrating Projet Volterra and the HGV (Jackson et al. 2009) and they acknowledged that there were still many limits to the automatic integration of humanities datasets. They proposed building systems that made used of *human* annotations in helping to link up diverse data sets in a meaningful way:

> Our investigations led us to conclude that there is a need among at least some humanities researchers for tools supporting collaborative processes that involve access to and use of complex, diverse and geographically distributed data resources, including both automated processing and human manipulation, in environments where research groups (in the form of "virtual organisations"), research data and research outputs may all cross institutional boundaries and be subject to different, autonomous management regimes (Hedges 2009).

The need for infrastructural solutions that deal not just with the complexities of humanities data but also with the fact that it is often geographically distributed and can belong to different organizations with different data management practices will be further explored in the next section that examines the requirements of "general" and "domain specific" humanities infrastructures.

"General" Humanities Infrastructures, Domain-Specific Needs, and the Research Needs of Humanists

To return to the idea of general requirements for a humanities cyberinfrastructure, the ACLS report concluded that at a bare minimum it will have to be a *public good, sustainable, collaborative, interoperable, and support experimentation* (ACLS 2006). These ideas have also been supported in a variety of other recent research on how to build a general humanities cyberinfrastructure. Franciska de Jong in a recent address to the European Chapter of the Association for Computational Linguistics (EACL) delineated similar requirements for an infrastructure that would support new "knowledge-driven workflows." This list included the "coordination of coherent platforms (both local and international)" in order to support the interaction of communities and the exchange of expertise, tools, experience and guidelines; "infrastructural facilities" to support both researchers and NLP tool developers (citing CLARIN as a good example); open access sources and standards; metadata schemata; best practices, exchanges, protocols and tools, and service centers that would be able to support heavy

computational processing (de Jong 2009). Her requirements go beyond those of the ACLS by also specifying several other important features: the need for a number of pilot projects between NLP researchers and humanists to test specific features of the infrastructure, flexible user interfaces that meet a variety of scholarly needs, and realistic evaluation frameworks that assess how well user needs are being met by all the components of the infrastructure.

Questions of general infrastructure were also considered at a 2007 international workshop that was hosted by JISC and the NSF. This workshop produced a report that explored how to build an infrastructure that would support cyberscholarship across the disciplines. They emphasized a number of necessary conditions for infrastructure, including: new methods for data capture, management and preservation of digital content, coordination at the national and international level, interdisciplinary research, and most importantly, digital content that is truly "open," or in other words, available for computational processing and reuse. The authors of this report also caution, however, that creators of cyberinfrastructure will need to understand that a single approach will not work for all disciplines while at the same time resisting the assumption that there are no standardized services to be offered across disciplines (Arms and Larsen 2007). A similar warning was given by the CSHE report on scholarly communication. "Although robust infrastructures are needed locally and beyond, " Harley et al. concluded, "the sheer diversity of scholars' needs across the disciplines and the rapid evolution of the technologies themselves means that one-size-fits-all solutions will almost always fall short" (Harley et al. 2010).

Specific advice in terms of designing VREs or infrastructures that can be widely adopted across disciplines has also been given by Voss and Procter (2009). "Creating an integrated e-research experience fundamentally relies on the creation of communities of service providers, tool builders and researchers working together to develop specific support for research tasks," Voss and Procter argued, "as well as the creation of a technical and organisational platform for integrating these tools into an overall research process." While they argued that interdisciplinary approaches must be investigated, they also stated that any infrastructure that is developed must address the fact that social or organizational/disciplinary behaviors and technological issues are closely related.

The ability to create an infrastructure that is both general enough to encourage wide-scale adoption and use and that can meet the various needs of different disciplines at the same time is a complicated undertaking. While lauding the ACLS report in general, Stuart Dunn also warned that:

> …the "not only discipline-specific" aspect of cyber infrastructure expresses both its strongest appeal and its main drawback: while generating new knowledge by working across and beyond established intellectual disciplines is at the heart of "digital scholarship", the lack of a disciplinary focus with which scholars can identify is another reason why the term VRE has not established itself (Dunn 2009).

While as Dunn observes interdisciplinary research is at the "heart" of digital scholarship, the lack of a disciplinary focus for infrastructures can make it hard for researchers to identify them as useful for their needs and has thus limited the uptake of potential tools such as virtual research environments.

Tobias Blanke has also explored how e-Science tools and methodologies (including virtual research environments) may or may not be able transform "digital humanities" into "humanities e-Science" (Blanke 2010). One of the most successful tasks of digital humanities Blanke noted is using sophisticated text encoding with markup such as that of the TEI to support text exchange between individual scholars or projects, but he also cautioned that it remained to be seen whether TEI could be similarly useful for text exchange between computational agents. Blanke uses this example to illustrate how the ways in which technologies have been used in the digital humanities will not always work for e-Science processes. One of the greatest challenges, Blanke asserted, will be to "use the experience gained in Digital Humanities to build integrated research infrastructures for humanities" (Blanke 2010). Building this integrated research infrastructure nonetheless also starts with examining in detail the *research workflows* of individual humanities *disciplines* and their specific needs according to Blanke.

The challenge of moving beyond the small ad-hoc projects commonly found in the digital humanities to more systematic research that can deliver specific pieces of a larger arts and humanities infrastructure is a frequently cited problem. A related issue is how develop a digital infrastructure that respects the individual questions and

research needs of specific disciplines while also working towards more general-purpose solutions. These questions were addressed by Blanke et al. (2008) who reported on a number of grass-roots initiatives within the U.K. and Germany that allowed them to form successful *partnerships* with science to address common problems and to adopt *new viewpoints* on old questions in humanities research. While the authors argued that large-scale infrastructure in e-Humanities would be useful "mainly in the provision of data and computational resources," they also insisted that research should neither avoid creating *local* solutions if necessary nor try making *universal* claims. They ultimately put forward the idea of "lean grids" and claimed that "a generic solution covering all research domains is likely to fail." The authors define "lean grids" as "approaches which incorporate the idea behind grids to share resources while at the same time not relying on building heavy infrastructures" (Blanke et al. 2008).

While Blanke et al. (2008) did see some utility in projects such as DARIAH, they also countered that the actual number of humanities and arts researchers who need grid computing is fairly small and stated furthermore that much of this research is conducted at smaller institutions that would lack the technical support necessary to use the grid even if it were available. The strongest reason they give, however, against *solely* developing computational solutions for humanities computing that rely on grid technology, is that:

> . ..most of digital research in the arts and humanities is done in an interactive manner as a way of humans requesting resources, annotating data or running small supporting tools. The model of 'running jobs' on the grid is alien to such research practices. At the moment at least, large and in particular grid-based infrastructures do not support interactive behaviour well (Blanke et al. 2008).

The insight that too exclusive a focus on grid computing might be counterproductive to most "conventional" types of digital humanities research is an important thing to consider when designing a potential infrastructure for humanists.

Understanding the actual research methods and processes of humanists is thus both an important and necessary step in building any kind of resources, tools, or infrastructure that will meet their needs. Any infrastructure that is developed without an understanding of the specific research questions and types of tools that are used by humanists in their daily research is unlikely to be successful. Indeed, the LaQuAT project emphasized this point in terms of linking up humanities databases:

> All linking-up of research databases needs to be based on a detailed understanding of what researchers could do or would want to do with such linked databases. In the future, one would need to investigate more realistic user scenarios and complex queries. More generally, there is a need to study the workflows currently used by researchers and understand how an infrastructure would contribute to or alter these (Jackson et al. 2009).

Yet the risks of developing tools or services that are too disciplinary-specific must also be considered, as illustrated by a recent discussion of e-science, the humanities and digital classics:

> We need disciplinary centers: classicists, for example, have their own specialized needs that involve the languages on which they focus. At the same time, we cannot have a flat organization, with each discipline managing its own infrastructure. A relatively large humanities discipline such as classics might be able to support its own unique systems, but that would only condemn us to an underfunded infrastructure that we could not sustain over time (Crane, Babeu and Bamman 2007).

These authors argue that while there will always be some need for specific disciplinary centers and projects, digital humanities developers must also be careful not to *only* develop isolated tools or systems that cannot "plug-in" to a larger system or infrastructure.

Nonetheless, Geoffrey Rockwell has argued that highly specific tool development has a larger place within cyberinfrastructure for the humanities, since many digital tools are *reinvented* as sources in the humanities are continuously *reinterpreted*:

> Tools are not used to extract meaning according to objective principles. In the humanities we reinvent ways of making meaning within traditions. We are in the maintenance by reinvention and reinterpretation business and we don't want our methods and tools to become invisible as they are part of the research. To shift tool development from researchers to infrastructure providers is to direct the attention of humanities research away and to surrender some of the research independence we value. To shift the boundary that defines what is legitimate research and what isn't is something humanists should care passionately about and resist where it constrains inquiry (Blackwell 2010).

While he understood the frustration of funders with the "plodding iterative ways of the humanities," Rockwell concluded that rather than suspending the development of new digital tools that humanists (digital or otherwise) would need to do a better job at "explaining the value of interpretation."

Beyond the viability of common digital tools, Brown and Greengrass (2010) have also argued that a single monolithic repository, VRE or portal structure that isn't designed to support *customization* is likely to meet with failure:

> The breadth of resources required to service the needs of such a heterogeneous community is unlikely to be encompassed by any single repository, or even a small cluster of major repositories. An access portal therefore needs to be 'customisable' to create links to and feeds from valued and commonly used sources (Brown and Greengrass 2010)

As with infrastructure for digital classics, while there are certainly common problems to be solved and tools to be developed that will work across humanities disciplines, there will still likely always be some need to *customize* tools and resources for different disciplines. Indeed, a recently released report for the ARL regarding the development of services for digital repositories also concluded that research libraries would need to attend to the "demand side" or the specific needs of different disciplinary user groups, for "digital repositories are as much about users as they are about content, so the development of high-value repository services requires understanding user needs and capabilities" (ARL 2009b). The report went even further and urged that "rather than developing technologies and hoping they will be usefully applied, libraries need more data, and discipline-specific data, on how a wide range of service consumers — institutions, libraries, scholars, and researchers — value services and want to use content."[579] Thus the importance of understanding the *specific* needs, research methods, and information habits of the different humanities disciplines will be an essential part of designing any larger humanities cyberinfrastructure.

This type of user modeling work is currently being conducted by the DARIAH project as explained by (Benardou et al. 2010a). As part of the preparation stage of DARIAH, a "conceptual model for scholarly research activity" is being created that is based on cultural-historical activity theory and is being expressed using the CIDOC-CRM. A two-pronged research program is being conducted by the Digital Curation Unit-IMIS of the Athena Research Centre that includes 1) an "empirical study of scholarly work" that will be based on the transcription and conceptual encoding of interviews with scholars (23 Europeans arts and humanities researchers) that can be considered "mainstream" users of digital resources and 2) the creation of a "scholarly research activity model" that is based on an *event-centric* approach that will be used to *formalize* the results of the empirical study into an actual systems model. After reviewing the extensive body of literature regarding scholarly information behavior and identifying a number of common processes across them (e.g. citation chaining, browsing, gathering, reading, searching, verifying), Benardou et al. proposed that one issue in terms of their own work was that all of these earlier studies present models that "view information behavior primarily as process; consequently, the world of information objects, data and documents, remains in them as a rule implicit" (Benardou et al. 2010a).

A number of other limitations were also identified Benardou et al. with the current literature regarding scholarly research activity, including that it 1) focused predominantly on the practice of "information seeking" rather than the whole "lifecycle of scholarly information use," a point also made earlier by Toms and O'Brien (2008); 2) concentrated primarily on the use of scholarly objects such as research publications from a library perspective and only "implicitly" on the use of *primary evidence* and secondary archives (a major research component of many humanities disciplines"; 3) privileged the information-seeking process over "object modeling"; 4) delineated a broad number of research activities and processes (e.g. scholarly primitives) but never attempted to *formally* define these entities or the relationships between them into a model of the research process; and 5) typically never went beyond "explanatory schematizations." For these reasons, Benardou et al. stated that their work objective would thus be to create a formal schematic of the research process, or basically to:

---

[579] Research into data curation and digital repositories by Martinez-Uribe and Macdonald (2009), where the authors interviewed life science researchers regarding their research practices and their likelihood of using a shared and open data repository, also stressed the importance of investigating actual user requirements for a digital repository and emphasized that *disciplinary* differences existed in terms of a desire for open data or digital archiving.

…establish a conceptually sound, pertinent with regard to actual scholarly practice, and elegant model of scholarly research activity, encompassing both "object" (structure) and "process/practice" (functional) perspectives, and amenable to operationalisation as a tool for: • structuring and analysing the outcomes of evidence-based research on scholarly practice and requirements and • producing clear and pertinent information requirements, and specifications of architecture, tools and services for scholarly research in a digital environment (Benardou et al. 2010a).

As a basis for their model, Benardou et al. used cultural-historical activity theory as developed by Leont'ev that has as its key concept, *activity*, or purposeful interactions of a subject with the world. Using the activity theory framework, they defined scholarly research as a "purposeful process" that is carried out by *actors* (whether individuals or groups) using specific *methods*. Research processes were broken into simpler *tasks* each of which could then be operationalized into specific *procedures*. Benardou et al. (2010a) also cautioned however that these research processes and their corresponding procedures should be considered to have a *normative* character and "convey what is believed by a community of practitioners to be good practice at any given time."

The CIDOC-CRM ontology was chosen to formalize their model of the research process and three key entities were defined including physical objects (e.g. objects that are found, stored and analyzed), conceptual objects (e.g. concepts that have been created, logical propositions) and information objects (e.g. conceptual objects that have "corresponding physical information carriers.") In sum, Benardou et al. explained that "the information objects are the contents of digital repositories; the physical objects are the original domain material; and the conceptual objects are the content of scientific theories." Their current research, however, focused exclusively on the relationship between information and conceptual objects. Another major entity that was defined was *Research Activity* and this was used as the basic "construct for representing research processes." This entity is typically associated with other entities such as *Procedure* that is in turn related to the *Methods* that are employed and the *Tools* or *Services* it requires. A special *Proposition* entity was also created that represents all hypotheses that are formulated or arguments that are made.

As was discussed earlier in this report, services form a key component of cyberinfrastructure, and indeed Benardou et al. reported that *services* formed an essential part of their ontological model of scholarly research. "*Services* thus become an important mediator between methods, procedures and information repositories," they explained, "From a functional perspective, affordances of digital scholarship are embodied in services available. From a teleological and methodological perspective, services evolve to better meet requirements" (Benardou et al. 2010a). The authors also concluded that both their own empirical study plus those in their extensive literature review have provided the "necessary substantiation on primitives" that along with an "elaboration of research goals" enable the development of a research model that is specific enough to develop appropriate digital services. The next stages of their research will be to operationalize this model and to tag all of the scholarly interview transcripts in terms of this model to validate its soundness. [580]

While the majority of research considered in this review stressed that developers should conduct needs assessment before designing tools and should also carry out user testing of prototypes or final products, Cohen et al. (2009) have also pointed out that such testing might not always necessarily yield the desired results. They suggested that it remained an open question whether "researchers or content communities can accurately assess what they need ahead of time, or whether they are biased toward repeating modes and interfaces they have already seen but which will be less effective at digital scholarship than more innovative software." The ways in which scholars current understanding of technology and comfort levels with certain types of interfaces may predispose them against new methodologies is a point worth considering in the design of any infrastructure.

VREs in the Humanities: A Way of Addressing Domain Specific Needs?

A variety of research has emphasized the development of virtual research environments or VREs for the humanities as one potentially useful building block for larger cyberinfrastructure. [581] Blanke (2010) promoted the idea of a humanities VRE that "would bring together several Digital Humanities applications into an

[580] The initial results of this work interviewing scholars and tagging transcripts has recently been published (Benardou et al. 2010b).
[581] JISC has recently released an extensive study that explores the role of VREs internationally in supporting collaborative research both within and across disciplines (Carusi and Reimer 2010).

integrated infrastructure to support the complete life cycle of humanities research" (Blanke 2010). One useful definition of VREs has been offered by Michael Fraser:

> Virtual research environments (VREs), as one hopes the name suggests, comprise digital infrastructure and services which enable research to take place. The idea of a VRE, which in this context includes cyberinfrastructure and e-infrastructure, arises from and remains intrinsically linked with, the development of e-science. The VRE helps to broaden the popular definition of e-science from grid-based distributed computing for scientists with huge amounts of data to the development of online tools, content, and middleware within a coherent framework for all disciplines and all types of research (Fraser 2005).

Fraser suggested looking at VREs as a one component of a digital infrastructure, rather than as stand-alone software, into which you could plug tools and resources. In fact, he argued that in some ways the terms VRE and cyberinfrastructure could almost be synonymous, with the one difference being that: "the VRE presents a holistic view of the context in which research takes place whereas e-infrastructure focuses on the core, shared services over which the VRE is expected to operate" (Fraser 2005). Fraser also stated that VREs are intended in general to be both collaborative and multidisciplinary.

On the other hand, Voss and Procter have recently criticized a number of VRE projects due to their either overly specific or generically designed architectures and an overall lack of interoperability. "VREs that have been built to date tend to be either specific configurations for particular research projects or systems serving very generic functions," Voss and Procter concluded, "The technologies used to build VREs also differ widely, leading to significant fragmentation and lack of interoperability"(Voss and Procter 2009). In order to promote greater interoperability, Voss and Procter suggested identifying common features that would be required from a generic infrastructure across disciplines by exploring the research lifecycle. They identified the following functionalities as common to research environments for almost all disciplines: authenticate, communicate and collaborate, transfer data, configure a resource, invoke computation, re-use data, give credit, archive output, publish outputs (formally and informally), discover resources, monitor resources, maintain awareness, data provenance, authentication and authorization. In addition, Voss and Procter also identified a number of research challenges that would need to be further investigated before the successful deployment of VREs. These challenges included understanding the factors that influence the adoption of VREs, learning how they are used differently in various disciplines, and considering their implications for scholarly communications.

The project of integrating eSAD with the VRE-SDM (VRE for the Study of Documents and Manuscripts) provides a useful example of a possible VRE for Classics. eSAD had independently developed a number of image processing algorithms for scholars working with ancient documents, and these were consequently "offered as functionalities wrapped in one or several web-services and presented to the user in a portlet in the VRE-SDM application" (Wallom et al. 2009). The authors also reported that eSAD was developing a knowledge base that could also be implemented in the same portlet or another one. Before these algorithms were implemented in the VRE-SDM they were difficult for researchers to access and required more processing power than many single systems possessed. Although developing an interface to connect the portal with the National Grid Service (NSG) ran into some complications, Wallom et al. concluded that this project had been a success for they had managed to create a "build and installation configuration toolkit" that could be used by the eSAD researchers to distribute the algorithms they had created onto the NGS.

The VRE-SDM project originally grew out of the "Building a VRE for the Humanities" or BVREH project [582] at Oxford University that was completed in September 2006. This project surveyed the use of digital technologies through a scoping study at Oxford and wanted to create set of priorities for developing infrastructure for humanities research. The researchers conducted a series of semi-structured interviews with humanities professors with the end goal being the creation of a set of scenarios describing typical researchers (Pybus and Kirkham 2009). The surveyed revealed a number of insights and demonstrated that:

> …the overall priorities of most interviewees concerned central hosting and curation of the digital components of research projects; and potential for a VRE to facilitate communications. The latter included the dissemination of results from projects, event notification, registers of research interests, collaboration beyond institutional boundaries, and the promotion of humanities research at Oxford more

[582] http://bvreh.humanities.ox.ac.uk/

generally. The cross-searching of distributed databases, project management tools, and directory services for hardware, software and other types of electronic resources were also noted as important requirements (Fraser 2005).

The researchers found that focusing on humanities professors current research and asking them to describe a "day in the life" of their work was a very effective approach that not only gave professors a sense of ownership in the process but also allowed documented user needs rather than technology to determine what demonstrators they developed.

After the survey was completed, the BVREH team used these common themes as the basis of a workshop for which they developed four demonstrators as standards-compliant portlets. One of the demonstrators created was the "Virtual Workspace for the Study of Ancient Documents (VWSAW)" that as stated above evolved into the VRE-SDM. The research of the BVREH project that built on this preliminary survey illustrated how conducting an analysis of the humanities research environment allowed them to map "existing research tools to specific components of the research life cycle" (Fraser 2005). Although technical development was not an end goal of this project, the survey still allowed them to make informed decisions regarding future infrastructure choices. While the BVREH team reported that *iterative* development was very important, their most important recommendation to others was to have VRE developers attend *meetings* with their intended users so they can come to understand both the type of research their users conduct and the kind of materials that are used (Bowman et al. 2010).

The experience of the BVREH project confirms the advice given by Voss and Procter (2009) in terms of building a VRE or that infrastructure designers need to consider the research methods of their intended users as well as their larger social and organizational context. "As the name virtual research environment implies," Voss and Procter explain, "the aim is not to build single, monolithic systems but rather socio-technical configurations of different tools that can be assembled to suit the researchers' needs without much effort, working within organisational, community and wider societal context" (Voss and Procter 2009). In other words, creators of infrastructure should not seek to build a universal monolithic system that can meet all possible needs but instead design extendable configurations of both general and domain specific resources and tools that users can adapt to meet their own research needs across the disciplines. While technological challenges remain, the sociological considerations of infrastructure decisions also need to be considered.

Despite the success of the VRE-SDM project, Stuart Dunn has cautioned that implementing VREs in the arts and humanities will be always be more complicated than in the sciences due to the "fuzzy" nature of research practices in the humanities (Dunn 2009). He argued that Google Earth, [583] although a commonly used tool and considered by some to be a sample VRE, should instead be looked at as a *component* of a VRE. A successful VRE in the humanities, Dunn stated, will have to meet number of requirements, including support for authentication so that scholars can record their methodology or publish how they researched their conclusions or created visualizations. In addition, users must have control or at the very least knowledge of how or if their data "is preserved, accessed and stored." Finally, a VRE for the humanities must have a clearly defined research purpose. Dunn ultimately proposed that Pleiades, rather than Google Earth, could be considered as a sample humanities VRE, since "it has a definable presence, it allows only authenticated users to contribute to its knowledge base, and there is a quantifiable and coherent research purpose" (Dunn 2009).

One other topic that Dunn brings up that was also a major concern documented throughout this research was the fact that digital publication is an act of *interpretative* scholarship that needs to make its data, methodology and decisions transparent and must create stable and citable results that can be verified and ideally tested and reused. Another overriding theme also illustrated by Dunn was the importance of designing technology that can support traditional or *existing* research practices as well as enable groundbreaking work:

The successful development and deployment of a VRE in the humanities or arts is contingent on recognizing that workflows are not scientific objects in their own right. Workflows in these disciplines are highly individual, often informal, and cannot be easily shared or reproduced. The focus of VREs in the arts and humanities should therefore be on supporting existing research practice, rather than seeking to revolutionize it **(**Dunn 2009).

[583] http://earth.google.com/

This conclusion that VREs, or indeed any cyberinfrastructure, will not be able to be used for *innovative* research until designers better understand how they can support *standard* research has been seen throughout this review. "Until analytical tools and services are more sophisticated, robust, transparent, and easy to use for the motivated humanities researcher," Borgman asserted, "it will be difficult to attract a broad base of interest within the humanities community" (Borgman 2009). As Borgman convincingly argues, if the digital tools and collections already available cannot be easily used, making an argument for the greater uptake of digital humanities research will be very difficult indeed.

New Models of Scholarly Collaboration

The need for more collaboration between humanities scholars both *within* the humanities and with *other* disciplines was called for repeatedly throughout the literature on both digital classics and humanities cyberinfrastructure. In their discussion of planning for cyberinfrastructure, Green and Roy (2008) concluded that unfortunately, "much of the daily activity of the humanities and social sciences is rooted in the assumption that research and publication form essentially an individual rather than a collaborative activity." While they were certain that the future of liberal arts scholarship would be in greater collaboration, they were uncertain if collaboration would be brought about by the creation of semantic tools that would make it easier to find and work with partners or if collaboration would be *forced* due to the need to organize increasingly huge amounts of data.

The CSHE report on scholarly communication also argued that greater collaboration will be important in cyberscholarship, but stressed as well the difficulties new collaborative models face, both disciplinary and financial:

> Collaborations around interdisciplinary grand challenge questions are especially complex, creating new demands for funding streams, administrative homes, sharing of resources, institutional recognition of individual scholars' contributions, and the need for participants to learn the "languages" of the multiple contributing disciplines (Harley et al. 2010, pg. 16).

Due to these various obstacles as well as a general resistance to change, the authors noted that there is still little *joint* authorship in the humanities. At the same time, Choudhury and Stinson hoped that the sheer scale of large digitization projects such as the Roman de la Rose would inspire humanists to pursue new collaborative forms of "data driven scholarship." Similar arguments were made by Stephen Nichols, who reflected that collaborative efforts were *required* to engage in the new types of scholarship that could now be conducted due to ever increasing amounts of data. "The typical digital project cannot be pursued, much less completed by the proverbial 'solitary scholar' familiar to us from the analogue research model," Nichols insisted, "Because of the way data is acquired and then scaled, digital research rests on a basis of collaboration at many levels" (Nichols 2009). Nichols listed three levels of collaboration that would become increasingly necessary: 1) partnerships between scholars and IT professionals; 2) new "dynamic" interactions between scholars both within the same discipline and across disciplines; and 3) collaboration between various IT professionals developing websites for scholars.

Despite many calls for collaborative scholarship, Borgman echoed the criticism of the CSHE report that is also found in (Crane, Seales, Terras 2009) about the continuing individualistic nature of much humanities scholarship. "While the digital humanities are increasingly collaborative," Borgman argued, "elsewhere in the humanities the image of the "lone scholar" spending months or years alone in dusty archives, followed years later by the completion of a dissertation or monograph, still obtains" (Borgman 2009). She agreed with an earlier insight of Amy Friedlander (Friedlander 2009) that in order to survive the digital humanities must move beyond large numbers of uncoordinated "boutique" projects into larger collaborative projects that can not only attract more funding but also ideally create larger sustainable platforms off of which more research can be built. The need to link up projects and researchers across disciplines and institutions has also been called for by the LaQuAT project:

> It is necessary to link up not only data, but also services and researchers — in the plural. Research in the humanities need no longer be an activity carried out by a single scholar, but rather by collaborating researchers interacting within an extended network of data resources,

digital repositories and libraries, tools and services, and other researchers, a shared environment that facilitates and sustains collaborative scholarly processes (Hedges 2009).

In addition to linking up data, services and researchers, Borgman also cited two other important issues for the success of digital humanities projects: 1) the need to move from a focus on *audience* to a focus on *participation* be it students, scholars, or the public and 2) the need to pursue collaborative relationships not only across the disciplines of the humanities but with computer scientists as well. Crane, Babeu and Bamman (2007) have also echoed this second point, stressing that humanists lack the resources and the expertise to go it alone in terms of developing infrastructure:

Unlike their colleagues in the sciences, however, humanists have relatively few resources with which to develop this new infrastructure. They must therefore systematically cultivate alliances with better-funded disciplines, learning how to build on emerging infrastructure from other disciplines and, where possible, contributing to the design of a cyberinfrastructure that serves all of academia, including the humanities (Crane, Babeu and Bamman 2007).

Scholars in the humanities thus need to learn to build relationships with colleagues in the sciences and to *repurpose* as many tools as possible for their own needs. Michael Fraser also made similar points in his recent piece on VREs:

For the most part, it is expected that computer science will act in partnership with other disciplines to lay the foundations, integrating methods and knowledge from the relevant subject areas. Humanities scholars, for example, cannot necessarily be expected to apply tools and processes (initially developed for the e-science community) effectively to their own subjects. Better to articulate the challenges and methods and sit down with the computer scientists. This is not an alien idea for many in the humanities - there is a long history of such partnerships (Fraser 2005).

The need of humanists to both outline their needs to computer scientists and to utilize their tools has also been made by Choudhury and Stinson (2007):

One of the imperatives for the humanities community is to define its own needs on a continuous basis and from that to create the specifications for and build many of its own tools….At the same time, it will be worthwhile to discover whether new cyberinfrastructure-related tools, services, and systems from one discipline can support scientists, engineers, social scientists, and humanists in others (Choudhury and Stinson 2007).

Partnerships between the humanities and computer science are thus not only necessary but also offer opportunities for truly interdisciplinary work. More research also needs to be conducted into how easily tools can be repurposed across disciplines.

On the other hand, scholars in computer science are also beginning to push harder for closer connections with the humanities. In an invited talk at the EACL, Franciscka de Jong called for rebuilding old liaisons between the natural language processing community and humanists:

A crucial condition for the revival of the common playground for NLP and the humanities is the availability of representatives of communities that could use the outcome, either in the development of services to their users or as end users. (de Jong 2009).

Toms and O'Brien have also recognized the need for greater collaboration between self-described e-humanists and the larger computer science community, noting that the lack of communication between these two groups had led to a limited awareness among e-humanists as to what tools were already available. "Perhaps due to the general and likely conditioned practice of not collaborating," Toms and O'Brien suggested, "they have not sought advice or collaborated with individuals from a host of other disciplines who have created tools and technologies that could support the humanities, e.g. information retrieval, natural language processing and linguistics to name a few" (Toms and O'Brien 2008). Both communities will thus have to begin making inroads to start building a common infrastructure.

Sustainable Preservation and Curation Infrastructures for Digital Humanities

Creating a cyberinfrastructure for the humanities involves not just creating new collaborative scholarly spaces for accessing distributed content and services but also for ensuring the long-term preservation, curation and sustainability of that content. Although the issue of digital preservation has received a great deal of attention in

the library community, [584] the specific challenges of preserving complicated digital humanities projects have received less attention. In her recent overview of this subject, Linda Cantara noted how most scholarship in this area seemed to assume that preservation was the responsibility of the *creator* of a digital project. At the same time, she observed most humanities scholars seemed to believe that institutional or digital repositories being created by research libraries would handle all the challenges of the creation of the metadata necessary for both preserving and maintaining the usability of digital content (Cantara 2006). The interim report of the "Blue Ribbon Task Force on Sustainable Digital Preservation and Access" also found that there was no agreement among stakeholders as to who should be preserving digital content or who should pay for it (NSF 2008).

This disconnect between content creators and preservers has only just begun to be addressed in the last few years. In 2009, a workshop entitled "Curriculum Development in Digital Humanities and Archival Studies" was held at the Archival Education and Research Institute and enabled digital humanists and archival researchers to meet and "collectively outline future directions for digital humanities research" (Buchanan 2010). Among the relevant issues discussed were the challenges of appraising collections of digital objects and designing navigable digital libraries that were accessible to non-experts, the importance of supporting collaboration, the continuing need for descriptive metadata to discover items on a granular level, the roles digital humanists have to play in constructing digital archives, and defining the skills needed by both archivists and digital humanists. "It is crucial that the digital humanities not only refine its extant disciplinary foci, but also begin to think generally and reflexively about its own sustainability, and that of its source data," Buchanan concluded, "As the digital humanities community continues growing in the direction of data collection and curation for born-digital (and not only paper-to-digital, or "digitized") materials, the field must begin to plan for regular surveys and monitoring of these valuable collections" (Buchanan 2010).

One of the largest-scale digital preservation research project that is currently underway is the EU funded Planets (Preservation and Long-Terms Access Through Networked Services). [585] According to the website, the Planets Project that began in 2007 has been created to "deliver a sustainable framework to enable long-term preservation of digital content, increasing Europe's ability to ensure access in perpetuity to its digital information." This work includes a number of deliverables such as providing preservation planning services, developing methodologies and tools for the characterization of digital objects, [586] creating "preservation actions" tools to "transform and emulate obsolete digital assets," building an interoperability framework to integrate diverse tools and services across a distributed network, providing a testbed to evaluate different preservation protocols, tests and services, and overseeing a dissemination program to promote vendor takeup and user training. Currently, they have developed a Preservation Planning Tool named Plato, [587] published a large number of white papers and research publications, and have released the initial Planets testbed. [588]

A smaller scale digital preservation organization that previously existed in the United Kingdom was the Arts and Humanities Data Service (AHDS) that was actively funded between 1996 and 2008 with the purpose or providing digital preservation and distributed access to digital humanities projects created in the United Kingdom. Funding for the AHDS, however, ended in 2008 and only the Archaeology Data Service survived. A great deal of the data, particularly that describing digital projects and ICT methodologies became part of the arts-humanities.net hub. Nonetheless, even when the AHDS still existed, Warwick et al. (2008b) argued that its ingestion processes for "completed" digital humanities projects as well as that of many institutional repositories was not sufficient for true long-term access to these resources:

---

[584] For example, see (Ball 2010) for an overview of the different tools that have been developed to assist institutional repositories in digital preservation and for an overview of digital preservation issues for libraries see (McGovern 2007).
[585] http://www.planets-project.eu/
[586] This research has also involved determining not just the significant properties of digital objects that must be maintained to ensure long-term meaningful access but also how these objects are used by different stakeholders and in what types of environments in order to determine appropriate preservation strategies (Knight and Pennock 2008).
[587] http://www.ifs.tuwien.ac.at/dp/plato
[588] https://testbed.planets-project.eu/testbed//. This testbed requires users to login, but also allows users to experiment with a number of different emulation and preservation services that can be tested against real data including external services that were not created by Planet, such as the popular JHOVE (JSTOR/Harvard Object Validation Environment) tool (http://hul.harvard.edu/jhove/).

The *de facto* solution is that individual institutions have become responsible for the electronic resources produced by their staff. However, although they may be willing to archive a static version of a resource in a repository and provide web server space, it is far more difficult for them to provide resources for active updating, since few institutional repositories have the expertise or personnel to ensure that the functionality of the resource is maintained. As a result, it seems likely that the slow decay of once functional digital resources will become more rather than less prevalent in future, at least in the case of the UK-based digital resources (Warwick et al. 2008b).

The authors also noted that older models of one-time deposit of digital data are of limited utility since data is rarely independent of its interface and the inability to update a resource once it has been deposited means that they quickly grow outdated and often unusable.

The inability of libraries and traditional repositories to maintain digital resources that are constantly increasingly in complexity as well as the digital tools that are needed to use them has also been described by Geoffrey Rockwell:

The more sophisticated digital works we create, the more there is that has to be maintained and maintained at much greater cost than just shelving a book and occasionally rebinding it. Centers and institutes get to the point that they can't do anything new because maintaining what they have done is consuming all their resources. One way to solve that problem is to convince libraries to take your digital editions, but many of us don't have libraries with the cyberinfrastructure. Another way to deal with this is to define certain tools as cyberinfrastructure so that they are understood as things that need ongoing support by organizations funded over the long-term. If the scale is right we might even have an economy of scale so that we could all pay for a common organization to maintain the commonwealth of infrastructure…. (Rockwell 2010).

Rockwell proposed that the Bamboo project might be an important first step in this direction since it is attempting to determine the common elements required for a digital research infrastructure across various disciplines and then plans to develop a consortium to build and sustain them in a distributed manner.

A recent report released by the ARL that explored the potential role of digital repository services in regards to research libraries has also made a number of recommendations to address these issues surrounding the systematic collection and preservation of digital projects and their long-term curation. This report observed that digital repositories have begun to develop quite rapidly and are quickly becoming a "key element of research cyberinfrastructure." In particular, the report proposed that research libraries should develop outreach strategies to researchers and scholars that have collected or created content that may have grown beyond their ability to manage: "Where these collections are of high value, local processes are needed to migrate early digital collections into an institutionally-managed service environment" (ARL 2009b). Research by Palmer et al. (2009) also identified the cataloging, collection and curation of digital materials, particularly the *personal* digital collections [589] of individual scholars, as important strategic services to be provided by research libraries in the future.

The report by the ARL also identified a number of other key issues that research libraries would need to address and services they would need to provide in order to develop digital repositories that functioned as part of a larger cyberinfrastructure. At the minimum, they proposed that digital repositories would need to offer the following services: long-term digital preservation, ongoing migration of content, access management, dissemination of research, metadata and format management, various types of discovery tools, digital publishing, and data mining or other forms of text analysis. Along with these essential services, the report also urged that research libraries should begin to develop services "around new content and old content in new forms."

While many initial repositories for research libraries had been originally conceived of as storage services for static PDFs of formally published faculty research output, the ARL report advocated that research libraries must plan for the fact that their institutions produce "large and ever-growing quantities of data, images, multimedia works, learning objects, and digital records" as well as recognize that "mass digitization has launched a new

[589] Palmer et al. (2009) have described these personal digital collections and their importance in great detail: "In the humanities, personal collections are the equivalent of finely curated special collections that have been expertly selected and controlled for quality and application. Rereading and notetaking are core functions with these collections. There is considerable potential for sharing and reuse of these collections, but the provenance and context of the materials from the scholar's research perspective is a large part of the value that would need to be retained and represented." (pg. 44) The CHSE study of faculty use of digital resources in teaching also noted the importance of personal digital collections (Harley et al. 2006a, Harley et al. 2006b).

scale of digital content collecting." This report also emphasized how digital repositories will need to be able to *integrate* the diverse digital content that already exists and that will continue to grow far outside of library-managed environments.

> Research practices will increasingly take advantage of strategies predicated on the availability of large amounts of widely accessible, rather than isolated and sparse, data. Many primary source materials supporting humanistic investigations — large corpora of texts, collections of images, and collections of cultural materials — will be complemented by many newly available and discoverable materials from disparate sources outside of library collections. To draw on content from these diverse sources, researchers will integrate use of library services and resources with funder-supported resources, commercially provided resources, and services and resources provided by other entities within the academy. Consequently, librarians will have much less control of the user experience than currently and will adopt more strategies that rely on collaboration with users. For instance, in areas such as curation and preservation of data, librarians will be regularly curating with, not just for, researchers (ARL 2009b).

The ARL report recognized that the mass amount of humanities materials that have been digitized both within open access projects and commercially licensed sources will be combined by researchers in new ways, and that they research library will need to develop new strategies to work with both its users and other academic units to support digital scholarship as well as preservation and curation of digital data. Abby Smith has also made similar arguments in her overview of how the research library will need to evolve in the 21st century to meet the needs of humanities scholars in particular:

> The accelerated development of digital humanities is an even more significant trend for research libraries, if only because humanists have been their primary clientele. Beyond the increasing use of quantitative research methods in the humanities, there is a growing demand by humanists to access and manipulate resources in digital form. With the primacy of "data-driven humanities," certain humanities disciplines will eventually grow their own domain-specific information specialists. While perhaps trained as librarians or archivists, such specialists will work embedded in a department or disciplinary research center (Smith 2008).

Many scholars and librarians have also stressed the need for research libraries to expand their preservation mission to include complicated digital objects and not just scholarly research publications.[590] Michael Fraser highlighted the importance of the ability of digital repositories to preserve more complicated content than just PDFs in the long-term infrastructure of VREs:

> Indeed, preserving the 'project', comprising data, publications, workflows and the 'grey' material of reports, notebooks and other forms of more nebulous communications is important in a research environment where much of the material is born and raised digital. The development of today's research by tomorrow's scholars depends on it (Fraser 2005).

Similarly, in their process of creating a digital library of the *Roman de la Rose*, Choudhury and Stinson also described the new *dual* responsibility of libraries to provide both physical and digital preservation:

> For while the curation of physical codices will remain an essential role for libraries, the collection and curation of digital objects will assume greater importance for libraries of the future, and the infrastructure, budgetary priorities, and strategic plans of library organizations would do well to account for this sooner rather than later (Choudhury and Stinson 2007).

A variety of research has also proposed that new *organizational* structures beyond the traditional research library or institutional repository (IR) may be needed to support such digital preservation and curation activities.[591] Sayeed Choudhury has recently described how Johns Hopkins University created its IR as "a "gateway" to the underlying digital archive that will support data curation as part of an evolving cyberinfrastructure featuring open, modular components" (Choudhury 2008). As was also argued by Abby Smith, they found that new roles were developing along with this new infrastructure including the development of "data humanists" or "data scientists."

A recent article in *Digital Humanities Quarterly* by W. A. Kretzschmar has also proposed that the only way to effectively support long-term and large scale humanities computing projects may be to find a "stable institutional setting" for them (Kretzschmar 2009). Crane, Babeu and Bamman (2007) have similarly argued that humanists may need to "develop new organizational structures to develop and maintain the services" on which they increasingly depend. The need to consider what type of organizations and funding will be required to

---

[590] See (Sennyey 2009) for a thorough overview of how academic libraries need to redefine their mission to meet the needs of digital scholarship.

[591] For example, the Digital Curation Centre (http://www.dcc.ac.uk/) in the United Kingdom has been established to provide advice and tools for the curation of research data and was first launched in 2004. Its website contains an extensive resources directory as well as a catalogue of digital curation tools.

maintain cyberinfrastructure in the humanities have also been explored by Geoffrey Rockwell. "When we look closely at civic infrastructure, we see that the physical infrastructure and service infrastructure are dependent on organizations for maintenance and operation," Rockwell explained, "In fact, if it is important that infrastructure last and be open, then the organization that maintains it is more important than the item itself"(Rockwell 2010). One major problem he listed was that funding bodies typically like to build *new* infrastructure rather than either budget for or fund its ongoing *maintenance*. He concluded that a realistic conception of infrastructure should include not just hardware components and "softer services" but also professionals that will be needed to operate and maintain it.

Green and Roy (2008) also maintained that new types of arrangements and institutions would be necessary in order to support and preserve digital scholarship. They provided an overview of several models, including privatization, the creation of open source models, or the creation of new types of "trans-institutional associations" such as HASTAC [592] that could reduce the risks that individual institutions need to take and start working towards the building of "discipline-based communities of practices." Whatever solution is chosen, they conclude:

> Although each of these models—privatization, open source, pay-to-have-a-say open source, and members-only or emergent transinstitutional associations—has its place in this emerging landscape, the key shift in thinking must be away from what can be done locally on an individual campus and toward how the campus can be connected to other campuses and how it can contribute to the refining of these new ways of doing scholarship and teaching (Green and Roy 2008).

In order to bring about this shift, they conclude with a number of important tasks that must be undertaken: 1) a strategic investment in cyberinfrastructure that will include a move from "collection development to content curation;" 2) the development and fostering of open access policies; 3) the promotion of "cooperation between the public and private sectors;" 4) the cultivation of leadership on all levels; 5) the encouragement of digital scholarship by creating national centers that support its growth; 6) the development and maintenance of "open standards and tools," and finally, 7) the creation digital collections that are both extensive and can be reused.

Questions of preservation and *sustainability* will likely remain difficult ones for the foreseeable future, and one useful listing of the components of sustainability as stated by Don Waters and recalled by Roger Bagnall are "a product people want, a functional governance structure, and a workable financial model" (Bagnall 2010). Similarly, the ARL report stated that the multi-institutional repository the HathiTrust [593] might find long-term success because they have already confronted the issue of "balancing governing and funding" (ARL 2009b). In fact, securing long-term funding, particularly for staff and basic technical infrastructure, is perhaps the critical issue that remains without many tractable solutions, as recognized by Edmond and Schreibman (2010):

> How do we ensure that the interfaces, web applications, or digital objects are maintained for future use by the scholarly community? We should be devoting more resources to the development of sustainable digital scholarship rather than accepting the fragility of the structures we create due to short-term or soft funding; resources that once created are situated outside the traditional funding and institutional structures in the humanities. Moreover we need to find long-term funding strategies for supporting the personnel and resources needed for these projects despite the fact that they are more typical of a science lab than a humanities project: programmers, servers, web developers, metadata specialists, to name but the most obvious (Edmond and Schreibman 2010).

In their own work with the Digital Humanities Observatory, they stated that that their organization was currently funded through 2011, with no clear business model or sustainability plan. Moving from "core funding" to piecemeal funding secured through grants they also noted frequently "diverts staff from "core activities" the infrastructure was designed to carry out." The authors also criticized the fact that almost exclusively project based funding had encouraged the creation of digital *silos* rather than integrated resources, a trend that projects such as Bamboo, CLARIN and DARIAH are seeking to address. Nonetheless, Edmond and Schreibman were not certain, as Peter Robinson has commented before, that such large infrastructure projects were necessarily going to be successful. "Generations of big projects, Europe's DARIAH and Project Bamboo not excepted," they reasoned, "seem to struggle with the notion that the right tools will turn the scholarly Sauls to Pauls, and

---

[592] http://www.hastac.org/

[593] The HathiTrust (http://www.hathitrust.org/) was originally created by the thirteen universities of the Committee on Institutional Cooperation and the University of California to establish a shared digital repository for preservation and access and includes all of the books digitized for these universities by Gogole Books. A number of other research libraries have also joined HathiTrust and currently the repository provides access to over 6 million volumes.

bring them in their droves into the digital fold. Others put forward the notion that generational change will bring us along regardless of our efforts for or against changes in modes of scholarly communication." But as was illustrated by the CSHE report (Harley et al. 206), longer term changes in acceptance and sustainability of digital scholarship will likely require far more than a simple changing of the guard.

Although many humanities scholars may feel that infrastructure and technical questions are the preserve of librarians and technologists, Neel Smith has convincingly argued that questions of openness, digital infrastructure and sustainability must be at the forefront of any humanist discussion of digital scholarship:

> Humanists can with some justification feel that the dizzying pace of development in information technology leaves them little time to reflect on its application to their area of expertise. And, after all, why should they concern themselves? As long as digital scholarship 'just works' for their purposes, isn't that enough? Here, as with software, the problem is that digital scholarship never 'just' works. The Homer Multitext project has focused on the choice of licences, and the design of data models, archival storage formats, and an architecture for network services because those decisions determine what forms our scholarly discourse can assume in a digital environment as definitively as code determines what a piece of software can accomplish (Smith 2010, pg. 136-137).

The Homer Multitext chose to use open data formats and free licenses so that their material could be both preserved and reused, two key components to sustainability. Similar arguments regarding the relationship between the ability to reuse materials and long-term sustainability have also been made by Hugh Cayless (Cayless 2010b). Cayless has recently proposed that studying how ancient texts have survived may provide us with some idea for how digital objects may be preserved. As Cayless explains, texts have typically been transmitted in four ways: accident, reuse through incorporation into other entities, republication or replication and durability of material (e.g. stone inscriptions). Through an overview of manuscript transmission and textual criticism, Cayless detailed the varying fortunes of Virgil, Sappho and the *Res Gestae* across the centuries. Cayless then touched upon a theme illustrated throughout this review, of how textual criticism needs to move beyond attempts to perfectly reconstruct the "original" text of an author by correcting the "errors" of scribes found in manuscripts (Bolter 1991, Dué and Ebbott 2009). "It is clear after centuries of studying the processes by which manuscripts are transmitted," Cayless argued, "that precise, mechanical copying was not typically the intent of those making new editions of classical works" (Cayless 2010b, pg. 144). Cayless stated that the methods of textual criticism would need to be adapted even further in terms of digital copies and their derivative formats.

Another pressing issue identified by Cayless was the fact that current digital rights management schemes rarely work well with digital preservation goals, for successful preservation requires the ability to distribute and migrate copies.[594] Cayless observed that Creative Commons licenses were one kind of license that dealt with varying levels of reuse or mashups. While he noted that one major frequently stated concern regarding sharing or making reuse available was that it would reduce an author's ability to profit from their own work, he also reported that some authors had stated that they had made more money by making at least part of their work freely available online. Cayless thus concluded that since making more of a work freely available increases the likelihood of it being quoted and reused, it also enhances its chances of being preserved:

> It seems therefore reasonable to argue that we have returned to a situation somewhat like the one that existed in the ancient world and furthermore that perhaps some of the processes that governed the survival of ancient works might pertain to digital media. As in ancient times, a work released into the electronic environment may be copied, quoted, reused or resold without the originator's having much control over what happens to it. There are legal frameworks for controlling what happens to copies of a work, but in practice they may be hard to apply or may not be worth the trouble. Some works may be licensed in such a way that there are no legal barriers to such treatment. What we have seen from the limited survey of ancient works above is that copying often provides the most promising avenue for long-term survival (Cayless 2010b, pg. 147).

While copying does tend to reflect the motivations of the current culture and is not without its complications, Cayless also pointed out that without copying, none of the texts of Sappho would have survived.

Another digital preservation issue that Cayless felt was misguided was a focus on preserving the "user experience," which he felt was typically defined as the *appearance* of the text on the page. "Modern printing

[594] This point was also made earlier by (Kansa 2007), and Cayless cites the LOCKSS (Lots of Copies Keeps Stuff Safe) program based at Stanford University (http://lockss.stanford.edu/lockss/Home) that includes an "open source, peer-to-peer, decentralized digital preservation infrastructure" as one potential model to be considered.

methods are completely unsuited to representing the appearance of ancient texts," Cayless insisted, "It wouldn't be possible to print a scroll on a modern laser printer without destroying its form. But there is absolutely no guarantee that the current standard form will be the dominant one in a hundred years" (Cayless 2010b, pg. 147). This concentration on the "user experience" Cayless proposed had created an overemphasis on *technology* rather than *content*. For the long-term preservation of content, Cayless suggested the use of text based markup technologies such as XML rather than document formats such as PDF:

> Text-based markup technologies, on the other hand, such as XML, allow for the presentation of documents to be abstracted out to a separate set of instructions. Instead of the document being embedded in the format, the format is applied to the document. In other words, the content becomes primary again, and the appearance secondary. This type of focus is very much in keeping with the ways in which ancient documents have reached us: none of their copyists would have argued that the text's appearance was as important as its content. The appearance will have changed every time the text was copied (Cayless 2010b, pg. 148).

Thus a renewed focus on *intellectual content* rather than *physical appearance* is not only important from a digital preservation perspective, Cayless argued, it is also more in keeping with how ancient texts were *transmitted*. In addition, Cayless stated that many ancient texts were transmitted with their commentaries, and that digital texts will need to be able to have their own modern versions of commentaries such as notes and annotations preserved as well. While both PDF and Microsoft Word have inflexible annotation models, Cayless noted that XML allows for more easy and flexible text annotation.

Finally, Cayless listed five interesting pieces of advice for digital archivists that offer food for thought for any long-term infrastructure planning for the humanities. First, as the future view or use of any work cannot easily be predicted, due care must be taken for preserving a large *variety* of digital resources. At the same time, Cayless also reiterated once again how "long-term survival may best be ensured by releasing copies from our control" (pg. 149). Second, as works have varying cycles of interest, long-term preservation must account for cycles of *disinterest* that could threaten the survival of a work. Cayless thus advised digital archivists to promote the use of their *whole* collection, including their lesser-known items. [595] Third, "self-sustaining communities of interest" may prove the most important factor in the long-term survival of works, so digital archivists should seek to help connect and facilitate communication between interested users and promote the growth of *communities*. Fourth, while an entire original object may not survive, the intellectual content might still be preserved (e.g. fragmentary texts or derivative works), so Cayless suggested that digital archivists should perhaps worry less about maintaining the integrity of digital objects outside of their curatorial control. Fifth, the more copies of a work that *exist*, the more likely it will be to survive, so digital archivists should extend efforts to obtain rights to *reproduce* digital resources without limitations.

Levels of Interoperability and Infrastructure

A key issue for any long-term preservation infrastructure will be the interoperability of its various components such as different digital repository platforms, diverse types of widely distributed content and heterogeneous data, and individual digital humanities applications, services and tools. The ARL report identified the pressing need of research libraries to increasingly engage with a "larger networked environment" and stressed that digital repositories could no longer be created as isolated collections or silos and instead needed to be designed "in ways that allow them to participate in higher-level, cross-repository services"(ARL 2009b).

In addition, the report by the ARL also asserted that by 2015 much technology that was once managed *locally* would instead be managed in a *distributed* and *virtualized* infrastructure such as through "cloud computing," [596]

---

[595] Similar arguments were made by Furuta and Audenaert (2010), who revealed that the *audience* for an original source is often the most widely forgotten actor and stated that: "Consequently, audiences have a significant, if indirect, hand in a work by determining what is accepted, what is copied, how it is packaged and which works survive."

[596] According to Webopedia (http://www.webopedia.com/TERM/c/cloud_computing.html), cloud computing is similar to grid computing, and it "relies on sharing computing resources rather than having local servers or personal devices to handle applications." Cloud computing applies supercomputing power to "perform tens of trillions of computations per second, in consumer-oriented applications" and accomplishes this by networking "large groups of servers, usually those with low-cost consumer PC technology, with specialized connections to spread data-processing chores across them. This shared IT infrastructure contains large pools of systems that are linked together." Virtualization technologies are frequently used to implement cloud computing. In terms of academic projects, Fedorazon has recently explored the use of cloud computing to support digital repositories using Amazon Web Services and Fedora (Flanders 2009).

either through collaboration both within and among institutions or through contracting to commercial providers. At the same time, they noted that library standards for interoperation would be increasingly overshadowed by more general network standards. As large datasets and other massive amounts of content become available, the report claimed that "research will grow more reliant on the production and use of large collections of data or primary source information organized into a plethora of repositories operating at national, disciplinary, and institutional levels" (ARL 2009b). Thus for researchers of the near future it will not be so important *where* a particular digital object or collection lives, but it will be essential that different repositories are able to *interact* with each other. As the ARL report concludes, "in this environment, interoperation between repositories and service technologies will be a pressing priority"(ARL 2009b). Gregory Crane has also argued this point recently, concluding that humanists will need networks of repositories and that the "greatest need is for networked repositories that can integrate collections and services distributed across the network in the short term and can then maintain these collections and services over decades" (Crane 2008).

The need for interoperable repository infrastructures has been addressed by Aschenbrenner et al. (2008), who have argued that for digital repositories to be successful they must become a natural and *integrated* part of users daily work environments, [597] this requires shared standards, description schemas and infrastructure, a task that they argued is beyond individual institutions and that will not be accomplished through the creation of "local repository islands:"

> Once it is possible to compose repository components through standard interfaces, repository managers can investigate which tasks they should take on themselves and which tasks to out-source to suitable service providers. This creates a much-needed competitive market where services can be plugged in with greater ease and for less cost…. However, for those new perspectives to manifest, the members of the repository community need to move closer together and develop a collaborative agenda (Aschenbrenner et al. 2008).

They also noted that many useful opportunities already existed in terms of promoting repository interoperability including scalable and on-demand generic storage infrastructures, large scale file-level processing for content analysis, more useful integration of application environments and a variety of light-weight preservation services such as "linking to community-wide format registries and migration services."

Romary and Armbruster (2009) have made similar arguments (with a focus on research publications) and concluded that digital repositories (thematic, geographical, or institutional) will only be successful and see major uptake if they are organized as large, central repositories (often organized by discipline such as Arxiv.org [598]) that can support faster dissemination, better services, and more effective preservation and digital curation. Centralized digital repositories, they also argued, will need to become part of the larger scholarly infrastructure that is developing:

> The specific vision that we have advocated in this paper goes into the direction of providing scientists with digital scholarly workbenches which, through a better coordination of technical infrastructures and adapted editorial support will provide both the quality and flexibility that is required for efficient scientific work. Even if we have focused here on the issue of publication repositories, which, for many reasons, lie currently at the centre of most debates, it is important to consider that this perspective is just one element within a larger set of digital scholarly services that have to be managed in a coordinated way (Romary and Armbruster 2009).

Although Romary and Armbruster have suggested that the creation of large-scale centralized digital repositories may be the best solution, the DRIVER [599] project has instead developed an infrastructure that supports federated access to over 249 individual digital repositories across Europe. Their initial research (Weenink et al. 2008, Feijen et al. 2007) identified a number of key issues that needed to be addressed to create such an infrastructure including intellectual property rights, data curation and long-term preservation. The DRIVER project guidelines mandated a standard way for repository data to be exposed but also provided technology to harvest "content from multiple repositories and manage its transformation into a common and uniform 'shared information space'" (Feijen et al. 2007). This shared information space provides a variety of services including 1) "services needed to maintain it" so data stores, indexes and aggregators are distributed on computers owned by various organizations; 2) the ability to add additional services as necessary; 3) a cleaning and enhancement service that

[597] Recent research by Catherine Marshall has offered initial analysis into scholarly writing and archiving practices through interviews with scientists in order to determine how to make the practice of scholarly archiving a more integral part of the research process (Marshall 2008).
[598] http://arxiv.org/
[599] http://www.driver-repository.eu/

standardizes content that is harvested into DRIVER records; 4) a search (SRW/CQL) and OAI-Publisher service that allows all DRIVER records to be used by external applications. Consequently any repository that wishes to participate can register within the DRIVER infrastructure and have their content "extracted, 'cleaned', and aggregated within an information space for integrated use" (Feijen et al. 2007). Ultimately, the DRIVER project focused on a centralized infrastructure with an extendable service model:

> Since the focus of DRIVER has been on developing infrastructure, it has not aimed to provide a pre-defined set of services. The infrastructure includes open, defined interfaces which allow any service providers working at a local, national or subject-based level, to build services on top. They will be able to reuse the data infrastructure (the Information Space) and the software infrastructure to build or enhance their systems. Services can therefore be developed according to the needs of users (Feiijen et al. 2007).

The DRIVER project illustrates the need and viability of a basic a common infrastructure for digital preservation and data storage, while also supporting the ability to develop innovative services by different projects.

One innovative approach to supporting even more sophisticated levels of repository interoperability has been introduced by Tarrant et al. (2009). Their work utilized the Object Reuse and Exchange (ORE) framework [600] that was developed by the OAI to support the "description and exchange of aggregations of Web resources" and was conducted as part of the JISC funded Preserv 2 project [601] that sought to find a way to replicate entire IRs across any repository platforms. As the OAI-ORE specification includes approaches for both describing digital objects and "facilitates access and ingest of these representations beyond the borders of hosting repositories," Tarrant et al. decided to see if it could be used to support a new level of cross-repository services. OAI-ORE was originally developed with the idea of creating descriptions of aggregations of digital objects (e.g. individual PDFs that are chapters of a book) and the relationships between them that could be utilized by any digital repository platform.

The OAI-ORE specification uses the concepts of *Aggregations* and *Aggregated Resources*, where an *Aggregation* represents a set of *Aggregated Resources*, with each resource and the *Aggregation* itself being represented by URIs. Tarrant et al. explained that in a sample OAI-ORE implementation the highest level *Aggregation* could be the repository itself and it could then contain various *Aggregated Resources* (e.g. research publications) each of which in turn could also contain their own *Aggregated Resources*. *Resource Maps* are used to describe individual *Aggregation*s (and can also link to only *one* aggregation) and each one must have a unique URI, but they can also make use of various namespaces and metadata schemas. OAI-ORE models can be represented in either RDF XML or the Atom syndication format.

The solution Tarrant et al. (2009) implemented made use of OAI-ORE with various extensions (e.g. writing export and import plug-ins, creating an individual application to create resource maps from digital objects stored in Fedora using RDFLib) [602] to replicate all digital objects in two different repositories (including their metadata and object history data) and enabled them to execute a lossless transfer of all digital objects between a Fedora and EPrints archive and vice versa. [603] By representing repository content through OAI-ORE Resource Maps, they proposed that many *different* digital repository platforms could then be used to provide access to the same content and this would enable an important "transformation from repositories-as-software to a services-based conception." OAI-ORE also specifies a number of different import and export interfaces that support both the exchange and reuse of digital objects, and Tarrant et al. (2009) believed that this feature could greatly aid in digital preservation by enabling simpler migration of objects between platforms:

> OAI-ORE provides another tool to help repository managers tackle the problem of long-term preservation, providing a simple model and protocol for expressing objects so they can be exchanged and re-used. In future we hope to see OAI-ORE being used at the lowest level within a repository, the storage layer. Binding objects in this manner would allow the construction of a layered repository where the core is the storage and binding and all other software and services sit on top of this layer (Tarrant et al. 2009).

---

[600] http://www.openarchives.org/ore/ and for more on the development of OAI-ORE, see (van de Sompel and Lagoze 2007).
[601] http://www.preserv.org.uk/
[602] A Python library for working with RDF, available at (http://rdflib.net/)
[603] Their approach also won the 2008 Common Repositories Interface Group (CRIG) challenge.

The use of OAI-ORE illustrates one potential data model that might be utilized by different digital classics or digital humanities repositories that wish to support the not just the *reuse* of their objects in various digital applications but their *replication* across different digital repository platforms.

In fact, OAI-ORE was used by Johns Hopkins University to represent various data aggregations in an astronomical data case study. Their data model included Resource Maps that represented Aggregations that contained multiple digital objects as well as objects *beyond* the individual Aggregation that were stored in a large number of different repositories. Choudhury argued that *individual* repositories could never be expected to include this level of data. "At some fundamental level, OAI-ORE acknowledges the realization that repositories are not an end," Choudhury remarked, "but rather a means to participate in a distributed network of content and services" (Choudhury 2008). Choudhury thus concluded that for institutional and consequently digital repositories to be successful they would need to define themselves as one *part* of the larger cyberinfrastructure, not as *the* infrastructure.

In addition to the complicated nature of repository interoperability an additional challenge for creating an interoperable infrastructure is that many individual digital humanities applications or tools are very specialized and do not support even limited interoperability. One means of addressing this problem according to Stephen Nichols would be if more digital projects were designed to be "tool-agnostic:"

> Rather than creating tools specifically for a given set of material, one can make platforms tool-agnostic: meaning simply that the site is designed to accommodate varied content. The capacity of a site to host multiple projects invites collaboration among scholarly groups who would otherwise each be putting up its own separate site. This in turn will promote scholarly communication and collaboration …in short, true interoperability. Technically such a model is not difficult to achieve; the problem lies elsewhere: in convincing scholars and IT professionals to think imaginatively and proactively by creating an 'ecumenical' platform for their original content, i.e. one that is general in its extent and application (Nichols 2009).

Cohen et al. (2009) made similar criticisms of the limited amount of tool interoperability, noting that despite the existence of various standards and methods that most tools have been built as "one-off, standalone web applications or pieces of software." Other issues they listed were the inability to import and export between tools and the difficulty of connecting different tools that perform the same task, citing as an example the vast proliferation of annotation software over the last five years (e.g. Co-Annotea, Pliny, Zotero, etc.). Furthering the problem is the fact that most digital tools only work with limited content collections, often because digital collections lack any way of communicating with tools such as an API.

In terms of creating larger VREs for the humanities, Voss and Procter have similarly argued that any successful VRE will need to include both generic and re-purposable components that can be used widely and believed that most research is indeed heading in that direction. "A wide range of commoditised components and systems are available," Voss and Proctor explained, "and efforts are underway to develop interoperability frameworks to foster flexible integration to form seamless collaborative work environments" (Voss and Procter 2009).

The VRE-SDM has followed this approach by making use of open source tools wherever possible (e.g. for annotation and for document viewing) and also made use of an existing VRE tool, the uPortal framework [604] (Bowman et al. 2010). They have used this framework in order to ensure interoperability with other VREs and virtual learning environments (VLE) for it allows them to reuse portlets from other projects and also makes their own components easier to reuse. This same framework can be customized by users who can compile their own interfaces using portlets that offer the tools and services they want. They are also currently using other newer standards (including Google gadget/OpenSocial) [605] to support the integration of various components. "This means that in the long-term the VRE will be able to provide tools to researchers across the humanities," Bowman et al. (2010) proposed, "Some, such as the viewing and annotation tools, will be relevant to the broadest range of scholars, while other more specialist tools can be added by individual users or groups as and when needed" (Bowman et al. 2010, pg. 96). As an example, they proposed that the VRE-SDM could deploy both its own tools and those of eSAD to explore collections such as Vindolanda and the LGPN. By using standards and tools that can be reused by various digital humanities projects, the VRE-SDM also hoped to

[604] http://www.jasig.org/uportal
[605] http://code.google.com/apis/opensocial

encourage other projects "to present their own tools and services for reuse within the environment." Their model of creating both a customizable and extendable architecture has also been followed by other projects such as TextGrid and DARIAH.

The varying challenges of interoperability (content, metadata, software, hardware, services, tools) are being addressed by all of the major digital humanities cyberinfrastructure projects including Bamboo, CLARIN, DARIAH, and TextGrid, all of which are discussed in further detail below, but a brief overview of their varying approaches to interoperability will be given here. The Bamboo project plans to develop a services platform that will both host and deliver shared services for arts and humanities research, and these services will run on the "cloud." The project also plans to adopt common standards and reuse other services and technology whenever possible (Kainz 2009). CLARIN is using Grid and Semantic Web technologies to ensure a full integration of services and resources and semantic interoperability respectively. Their ultimate plan is to create a "virtual, distributed research infrastructure" through a federation of trusted digital archives that will provide resources and tools (typically through web services) and provide users with a secure log-on (Váradi et al. 2008). Similarly, DARIAH is exploring the use of Fedora and the IRODS data grid technology to create a distributed research infrastructure that is secure, customizable and extendable (Blanke 2010).

A recent article by Aschenbrenner et al. (2010) has further examined how the DARIAH infrastructure will support the *federation* of various digital archives, an important task since research questions within the humanities often require materials that are stored in different locations. In order to support robust interaction between different agents in an open repository environment, they have broken down "interactions for repository federation" into three layers: physical, logical and conceptual. Similarly they have also identified six attributes of interoperability that must be addressed by any federated system: 1) digital object encoding (e.g. "byte serialization" for characters); 2) digital object syntax ("the strings and statements that can be used to express *semantics*" e.g. XML grammars), 3) semantics for digital objects, ("the meaning of terms and statements in a certain context", e.g. metadata formats, controlled vocabularies, or ontologies); 4) protocols (how different intellectual entities relate to one another within an information system, e.g. OAI-ORE, the METS standard [606]); 5) patterns ("identifies recurring design problems in information systems and present a well-proven generic approach for its solution, consisting of the constituent components, their responsibilities and relationships", e.g., harvesting through OAI-PMH), and 6) architectures ("specifies the overall structure, capabilities of and interactions between system components to achieve an overall goal") (Aschenbrenner et al. 2010).

In their approach to interoperability, or as they more specifically define it to promote larger, federated eHumanities systems, the creators of TextGrid have suggested the creation of federated semantic service *registries*, or registries that provide descriptions of the services and resources individual digital humanities projects or systems provide so that they can be discovered and potentially reused both by users and other systems. Aschenbrenner et al. (2009) posited that "standardized descriptions of services and other resources will be a precondition for building shared, federated registries" and will have the added benefit of "enabling a central query interface." Such a registry they also proposed will require a domain ontology that they have preliminarily developed. Aschenbrenner et al. (2009) also observed that in the last few years what they defined as "eHumanities Digital Ecosystems" have sprung up in great numbers. "The big challenge ahead," they thus concluded, "is now to see how these subsystems can begin to merge into one larger eHumanities DE while still maintaining their individual characters and strengths" (Aschenbrenner et al. 2009). Two essential prerequisites for such successful interoperability they argue are: "loosely coupled services" and the "visibility of resources." While they proposed a reference ontology for both services and documents in eHumanities, they also stressed that any infrastructure design must take user needs into account and ideally have users involved from the very beginning. "Novel infrastructure that is imposed on the user will fail," Aschenbrenner et al. advocated, "TextGrid has domain experts as core partners in the team, and these experts are shaping issues such as standards and community-building" (Aschenbrenner et al. 2009).

[606] For a useful overview of how METS and OAI-ORE differ and how they might be mapped in order to support greater levels of interoperability, see (McDonough 2009).

TextGrid thus made use of both domain experts and computer scientists in terms of defining standards for their project, and Aschenbrenner et al. reported that TextGrid has utilized nothing but open standards to promote the fullest amount of interoperability. They also took into account the three different interoperability layers identified by the European Information Framework: "technical, semantic and organizational" (Aschenbrenner et al. 2009). Earlier research by the TextGrid group had also highlighted the challenges of both *syntactic* and *semantic* differences in humanities data sets in terms of achieving meaningful data integration. "In the humanities, the major obstacle to data interoperability is syntactic and semantic heterogeneity," Dimitriadis et al. stated, "Roughly speaking, it is the differences in terminology that make it so difficult to cross the boundaries and create a joint domain of language resources that can be utilized seamlessly" (Dimitriadis et al. 2006). Similar research by Shen et al. (2008) had also reported that two major types of interoperability challenges for digital libraries were syntactic and semantic, with syntactic being at the level of *applications* and semantic interoperability as the "knowledge-level interoperability" that allows digital libraries to be integrated and includes "the ability to bridge semantic conflicts arising from differences in implicit meanings, perspectives, and assumptions, thus creating a semantically compatible information environment."

Although the development and use of standards to promote interoperability was called for by many projects such as TextGrid, other research including that by the LaQuAT project has also pointed out that standards have their *limits* as well:

> While there are a variety of standardisation activities with the aim of increasing interoperability between digital resources and enabling them to be used in combination, standardisation alone is unlikely to solve all problems related to linking up data. Humanists still have to deal with legacy data in diverse and often obsolete formats, and even when standards are used the sheer variety of data and research means that there is a great deal of flexibility in how the standards are applied. Moreover, standards are generally developed within particular disciplines or domains, whereas research is often inter-disciplinary, making use of varied materials, and incorporating data conforming to different standards. There will inevitably be diversity of representation when information is gathered together from different domains and for different purposes, and consequently there will always be a need to integrate this diversity (Hedges 2009).

Hedges argued that the realities of legacy data in the humanities, the differing application of standards, and the domain specificity of many standards necessitates the design of infrastructure solutions that can integrate diverse data. He suggested that research work by the grid community on the "integration of structured information" and turning data repositories into "virtualized data resources" on a grid may allow digital repositories to hide the "heterogeneity of digital objects" from their users, rather than trying to force all data into one standard. Whatever solutions are pursued this section has indicated that the question of interoperability exists on many different levels and will present significant challenges for any infrastructure design.

The Future of Digital Humanities and Digital Scholarship

Christine Borgman both started and ended her report on the future of digital humanities with five questions the community will need to consider in order to move forward as a discipline: "What are data? What are the infrastructure requirements? Where are the social studies of digital humanities? What is the humanities laboratory of the 21st century? What is the value proposition for digital humanities in an era of declining budgets?" (Borgman 2009).

The CSHE report also outlined a number of topics that deserve further attention in order to promote digital scholarship and new forms of scholarly communication. They recommended that tenure and promotion practices needed to become more nuanced along with a complementary reexamination of the processes of peer review. For scholarly communication to evolve, they concluded that business models also need to be developed that can sustain high quality and affordable journals and monographs. Additionally, they suggested that more sophisticated models of electronic publication are needed that can "accommodate arguments of varied length, rich media, and embedded links to data" (Harley et al. 2010). Finally, they addressed the importance of "support for managing and preserving new research methods and products including components of natural language processing, visualization, complex distributed databases, and GIS, among many other" (Harley et al. 2010).

The issue of the future of the digital humanities was also recently addressed by a THATCamp [607] that was held in Paris, France in May of 2010 [608] and the group issued a "Digital Humanities Manifesto" that included both a definition of digital humanities and general guidelines for ensuring a successful future for the digital humanities as a whole. These guidelines called for open access to data and metadata that must be both technically well-documented and interoperable, for greater and more open dissemination of research data, code, methods and findings, for the integration of digital humanities education within the larger social science and humanities curriculum (including formal degree programs in the digital humanities with concurrent promotion and career opportunities), for the definition and development of best practices that meet real disciplinary needs, and for the iterative creation of a scalable digital infrastructure that is based on real needs as identified by various disciplinary communities. The final section of this report will provide an overview of various large humanities cyberinfrastructure projects and how they are beginning to meet some of these needs.

Overview of Large Cyberinfrastructure Projects

The last five years has seen the creation of a large number of international and national cyberinfrastructure or e-science/e-research/e-humanities projects, creating a fragmented environment that has made it challenging to determine what type of infrastructure has already been built or may yet be created. Consequently a number of the most important projects, arts-humanities.net, [609] ADHO—Alliance of Digital Humanities Organizations, [610] CLARIN, [611] centerNET, [612] DARIAH, [613] NoC-Network of Expert Centres (in Great Britain and Ireland), [614] Project Bamboo [615], and TextGrid [616] formed a coalition named CHAIN [617] in October 2009. CHAIN, or the Coalition of Humanities and Arts Infrastructures and Networks plans to act as "a forum for areas of shared interest to its participants" including: advocating for strengthening digital infrastructure for the humanities, developing business models, promoting interoperability for resources, tools and services, promoting best practices and technical standards, developing a "shared service infrastructure," and widening the geographical scope of the current coalition. The various organizations had previously organized a panel at Digital Humanities 2009 to explore these same issues (Wynne et al. 2009) and met again at the 2010 Digital Humanities Conference.

This section will briefly provide an overview of each of these organizations as well as several other important national humanities infrastructure projects (Digital Humanities Observatory, DRIVER, TextVRE, SEASR) and the current work they have done.

Alliance of Digital Humanities Organizations (ADHO)

The Alliance of Digital Humanities Organizations (ADHO) is an "umbrella" organization "whose goals are to promote and support digital research and teaching across arts and humanities disciplines." [618] The organization was set up initially to more closely coordinate the activities of the Association for Computers in the Humanities (founded in 1973), the Association for Literary and Linguistic Computing (founded in 1978) and the Society for Digital Humanities/ Société pour l'étude des médias interactifs (founded in 1986). The ADHO is administered by a steering committee and membership in this organization can be obtained through subscription to *Literary &*

[607] THATCamp is an acronym for "The Humanities and Technology Camp" (http://thatcamp.org/), a project that was started by the Center for History and New Media (CHNM) at George Mason University. THATCamps have been held in various cities and have been described as "unconferences" "where humanists and technologists meet to work together for the common good." An unconference is one that is typically organized day-by-day by its participants according to their interests and is typically small, informal, short, inexpensive, informal, non-hierarchical according to the project website.
[608] http://www.digitalhumanities.cnrs.fr/wikis/tcp/index.php?title=Anglais
[609] http://www.arts-humanities.net/
[610] http://www.digitalhumanities.org/
[611] http://www.clarin.eu/
[612] http://www.digitalhumanities.org/centernet/
[613] http://www.dariah.eu/
[614] http://www.arts-humanities.net/noc/
[615] http://projectbamboo.org/
[616] http://www.textgrid.de/en.html
[617] http://www.dariah.eu/index.php?option=com_content&view=article&id=107:chain-dariah-participates-in-an-international-coalition-of-arts-and-humanities-infrastructure-initiatives&catid=3:dariah
[618] http://digitalhumanities.org/about

Linguistic Computing. The ADHO is also responsible for overseeing a number of digital humanities publications [619] including peer reviewed journals such as *Literary & Linguistic Computing*, *Digital Studies*, *Digital Humanities Quarterly*, *Computers in the Humanities Working Papers*, and *Text Technology*, a number of book series such as Blackwell's *Companion to Digital Humanities* and *Digital Literary Studies*. The ADHO website provides a list of community resources and hosts the Humanist discussion group archives. Finally, this organization also oversees the annual Digital Humanities Conference.

arts-humanities.net

The arts-humanities.net website provides an "online hub for research and teaching in the digital arts and humanities" that "enables members to locate information, promote their research and discuss ideas". This hub has been developed by the Centre for e-Research (CeRch) at King's College London (KCL) and is coordinated by Torsten Reimer. It incorporates several projects including "ICT Guides database of projects and methods" that was originally developed by the AHDS and the original arts-humanities.net developed by Reimer for the AHRC ICT Methods Network. Initial funding was provided by the AHRC and the project is now supported by JISC. A number of projects and groups contribute to this hub including the Network of Expert Centers, the Digital Humanities Observatory, the Oxford e-research center, and the Arts & Humanities e-Science Support Center (AHeSCC)

This hub has over 1400 registered users and by registering you can create a blog, participate in discussion forums, and tag various tools and projects. The entire website can be browsed by subject discipline, so for example, browsing by "Classics and Ancient History" [620] brings the user to a list of resources in the subject including recently added digital projects, events, additions to the research bibliography, calls for papers and blog entries (among others). Arts-humanities.net also contains several major components including a catalogue of digital resources, a directory of digital tools used in projects, a computational methods taxonomy, a research bibliography and library of case studies and briefing papers, an events calendar and list of calls for papers, a community forum that includes a discussion forum, a list of member's blogs and user groups, and an actively updated list of job postings.

The "Projects" section that serves as a "catalogue of digital scholarship" is one of the major resources on this website and provides several hundred detailed records on digital arts and humanities projects. [621] While the focus is on U.K. projects, the details provided are extensive including the types of digital resources created, the technical methods used, the data formats created, the types of tools used to create the resource and also several subject headings. The projects catalogue can be searched or browsed alphabetically, by method, discipline or content created. The actively updated digital tools section [622] can also either be searched by keyword or be browsed alphabetically (as well as by license, lifecycle stage, platform, subject tags (user created), and supported specifications. Another major component of the website is the ICT methods taxonomy [623] that includes seven broad method categories such as communication and collaboration or data analysis. Each method includes a list of sub-methods that leads to a description and a full list of projects using that method.

The arts-humanities.net hub fulfills a number of important requirements in terms of developing humanities infrastructure as outlined above. It is collaborative, supports the creation of communities of interest, and has created a central place to describe tools and methods to support best practices.

centerNET

centerNet is an "an international network of digital humanities centers formed for cooperative and collaborative action that will benefit digital humanities and allied fields in general, and centers as humanities

[619] http://www.digitalhumanities.org/publications
[620] http://www.arts-humanities.net/disciplines/classics_ancient_history
[621] http://www.arts-humanities.net/project
[622] http://www.arts-humanities.net/tools
[623] http://www.arts-humanities.net/ictguides/methods

cyberinfrastructure in particular." [624] This network grew out of a meeting hosted by the National Endowment for the Humanities (NEH) and the University of Maryland, College Park in 2007 and was created in response to the ACLS Report on Cyberinfrastructure in the Humanities (ACLS 2006). The largest component of this website is an international directory [625] of over 200 digital humanities organizations that can be viewed alphabetically. Each organization entry includes a small description and a link to their website and inclusion in the directory simply involves a request to join. As of July 2010, a new beta website [626] for centerNET was announced that currently includes a Google Map of all of the registered digital humanities organizations, an aggregated web feed from all registered centers websites or blogs, an updated directory of centers that can be searched or limited by geographic category, and a resources list that includes a link to a frequently updated "digital research tools wiki." [627]

CLARIN

CLARIN (Common Language Resources and Technology Infrastructure) is a pan-European project that is working to "establish an integrated and interoperable research infrastructure of language resources and its technology" it "aims at lifting the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and therefore enabling eHumanities." [628] The integrated environment of services and resources will be based on grid technologies, use Semantic web technologies to ensure semantic interoperability, and be extendable so that new resources and services can be added. CLARIN plans to help linguists improve their models and tools, aid humanities scholars in learning to access and use language technology, and to "lower thresholds to multilingual and multicultural content." CLARIN ultimately plans to build a "virtual distributed research infrastructure" through a "federation of trusted archive centers that will provide resources and tools through web services with a single sign-on" (Váradi et al. 2008)

CLARIN also seeks to address the issue of heterogeneous language resources in fragmented environment and seek particularly to connect resources and tools that already exist with scholars in the humanities:

> The benefits of computer enhanced language processing will become available only when a critical mass of coordinated effort is invested in building an enabling infrastructure, which will make the existing tools and resources readily accessible across a wide span of domains and provide the relevant training and advice (Váradi et al. 2008).

The CLARIN project especially wishes to turn the large number of existing and dispersed technologies and sources into "accessible and stable services" that users can share and repurpose. A number of other key themes they are dealing with include "persistent identifiers, component metadata, concept registries, and support for virtual collections" (Dallas and Doorn 2009).

As CLARIN is a very large project with many deliverables, its organization is divided into eight work packages: management and coordination, technical infrastructure, humanities overview, LRT overview, information gathering and dissemination, intellectual property rights and business models, and construction and exploitation agreements. CLARIN is still in its *preparatory* stage and envisions two later phases, a construction phase and an exploitation phase. This preparatory phase has a number of objectives including organizing the funding and governance in 22 countries and thoroughly exploring the technical dimension, for as Váradi et al. admit "a language resources and technology infrastructure is a novel concept." CLARIN is also fully investigating the user dimension, and are undertaking an analysis of how language technology is currently used in the humanities to make sure that all developed technical specifications meet the actual needs of humanities users. This scoping study and research includes undertaking a number of typical humanities research projects to help validate developed prototypes. In addition, they plan to conduct outreach to less technologically advanced sections of the humanities and social sciences to promote the potential use of language resources and technology (Váradi et

[624] http://digitalhumanities.org/centernet/
[625] http://digitalhumanities.org/centernet/?page_id=4
[626] http://digitalhumanities.org/centernet_new/
[627] http://digitalresearchtools.pbworks.com/
[628] http://www.clarin.eu/external/index.php?page=about-clarin&sub=0

al. 2008). CLARIN is also actively seeking to bring the humanities and language technology communities together and they thus plan to collaborate with DARIAH in this and other related areas.

One example of a humanities case study was reported by Villegas and Parra (2009), who explored the scenario of a social historian wishing to conduct a search of multiple newspaper archives. They found that providing access to primary source data that was "highly distributed and stored in different applications with different formats" was very difficult and that humanities researchers required the "integration and interoperability of distributed and heterogeneous research data." Villegas and Parra included a detailed analysis of the complicated steps required to create a final environment where the user could actually analyze the data. They also provided some insights for further CLARIN research and ongoing case studies, namely that: 1) humanists need to be made better aware of *existing* linguistics resources and tools; 2) users need integrated access to data with more *automated* processes to simplify laborious data gathering and integration tasks; 3) the use of standards and protocols would help make data integration easier; 4) NLP often tools need textual data to work on but many data providers do not have their data in a *textual* format; 5) the use of web services with standardized interfaces and strongly typed XML messages could help guarantee interoperability of resources and tools. Nonetheless, the authors also admit that this last desideratum will require a great deal of consensus among service providers.

In addition, to studying how humanities users might use language tools and resources, CLARIN plans to include language resources for all European languages in participating countries and it has currently defined BLARK, or a Basic Language Resources Toolkit, that will be required for each well-documented language. BLARKs must consist of two types of lexica, one "form based" and one "lexical semantic," or essentially a treebank and an automatically annotated larger corpus. As part of this work, CLARIN has recently made a number of services available on their website under their "Virtual Language Observatory." [629] Included among these services are massive language resource and language tool inventories that can be searched or browsed by faceted metadata.

Finally, the CLARIN website also provides access to newsletters, scientific publications, and extensive readable documentation [630] on the technological decisions of CLARIN. Documentation is available regarding the development of their concept registry, component metadata, persistent identifiers, long-term preservation, standards and best practices, text encoding, virtual collections, service oriented architecture, and web services interoperability.

DARIAH Project

DARIAH [631] stands for "Digital Research Infrastructure for the Arts and Humanities" and has been funded to "conceptualize and afterwards build a virtual bridge between different humanities and cultural heritage data resources in Europe" (Blanke 2010). The DARIAH project commenced in September 2009 and seeks to improve access to the hidden resources of archives, libraries and museums across Europe. DARIAH involves a consortium of over 14 partners from 10 countries. [632] Similar to CLARIN, the work for DARIAH has been divided into a number of work packages: Project Management, Dissemination, Strategic Work, Financial Work, Governance and Logistical Work, Legal Work, Technical Reference Architecture, and Conceptual Modeling.

Currently DARIAH plans to explore the use of the Fedora digital repository, the IRODS data grid technology developed by the San Diego Supercomputing center, and will be built on top of the "existing EGEE gLite infrastructure." [633] Tobias Blanke has offered a recent overview of DARIAH's basic architecture:

> DARIAH will integrate pan-European humanities research data collections by using advanced grid and digital repository technologies. Formally managed digital repositories for research data can provide an effective means of managing the complexity encountered in the humanities in general, and will take on a central and pivotal role in the research lifecycle. The DARIAH infrastructure will be fundamentally distributed and will provide arts and humanities publishers and researchers alike with a secure and customizable environment within which to collaborate effectively and purposefully (Blanke 2010).

[629] http://www.clarin.eu/vlo/ and for more details on this resource, see (Uytvanck et al. 2010).
[630] http://www.clarin.eu/external/index.php?page=publications&sub=3
[631] http://www.dariah.eu/
[632] http://www.dariah.eu/index.php?option=com_docman&task=doc_download&gid=301&Itemid=200
[633] http://technical.eu-egee.org/index.php?id=149

DARIAH thus seeks to build a secure, distributed and customizable infrastructure. For the next two years DARIAH will be in a preparatory phase and hopes to convince national funding bodies to transform DARIAH into a service.

Aschenbrenner et al. (2010) have stated that the major strategy behind the DARIAH repository infrastructure is that individual repositories should remain *independent* and evolve over time (thus remaining *open*) but at the same time all contents and tools provided through DARIAH should be appear to researchers as if they were using a *single* platform (thus a *closely-knit* infrastructure). Various institutions and researchers have been invited to contribute their content and tools to the DARIAH infrastructure, and Aschenbrenner et al. reported that they are committed to supporting *reasonable* levels of semantic diversity and interoperability:

> Linking diversity is at the core of Dariah's philosophy. Disciplines in the humanities differ greatly with regard to their resources – their data, tools and methodologies. Moreover, innovation is sometimes associated with introducing variations into their data, tools, or methodologies, thereby reinforcing heterogeneity even within a single discipline. Through linking this diversity Dariah aims to build bridges, to enable researchers from different disciplines or cultural backgrounds to collaborate on the same material in a common research environment, and to share their diverse perspectives and methodologies (Aschenbrenner et al. 2010).

DAR IAH seeks to both learn from and be interoperable with other repository federation initiatives such as DRIVER and Europeana [634] and consequently has also decided not to *enforce* rich metadata guidelines. The project, however, is still determining how to deal not just with diversity among research data in the humanities but among the different type of agents that will participate in the DARIAH environment (e.g. collaboration platforms, various service registries, private and public archives, content or service providers, applications and data created by individual scholars). Aschenbrenner et al. (2010) also emphasized that *interoperability* is not the same as *uniformity* so any interoperability scheme used for DARIAH may differ significantly from internal metadata schemes used at individual partner sites. Currently DARIAH is working on "enabling an evolutionary metadata approach" where scholars can start with minimum annotation and this can be extended over time by other scholars or automatically.

In order to support the federation of a wide variety of repositories and other agents, DARIAH has created a prototype federation architecture [635] for "exposing repository events" that is based on a "hybrid push/poll notification pattern" using Atom and supports "CRUD" [636]events:

> Since the Dariah environment will prospectively consist of a continuously growing number of agents that may depend on each other's state, a notification pattern is a suitable architectural building block. The Atom feeds containing the events of Dariah repositories can be consumed by decentralised agents, decentralised meaning that they only use the data available from the Atom feeds and do not plug into any proprietary repository interfaces. Such decentralised agents could hence be built by other initiatives consuming the event messages from various sources without the repositories being aware of them (Aschenbrenner et al. 2010).

DARIAH has tested this infrastructure by creating an experiment that linked TextGrid, an iRODS and a Fedeora test server into a single federation and both replicated digital objects across the different repositories and created an index of all the TEI/XML objects in the federation. Aschenbrenner et al. (2010) concluded that the use of Atom will not only "ensure coherence among decentralised agents" but as a *lightweight* protocol that is "deeply embedded into the web environment of HTTP-based, ReSTful Services" will serve as a gateway to a number of existing tools and improve the *scalability* of DARIAH as a whole. Some remaining challenges to be addressed by this infrastructure include user and rights management and the need for persistent identifiers for digital objects. [637]

One of the major projects of this first stage of DARIAH will be to build two demonstrators that prove this technical architecture is feasible. The first demonstrator will be "to construct an exemplary Service-Oriented Architecture to demonstrate the viability of the technologies identified in the technology review" and the

---

[634] http://www.europeana.eu/

[635] In addition to using the Atom syndication format (http://www.ietf.org/rfc/rfc4287.txt), this prototype made use of the Metadata Object Description Schema (MODS) (http://www.loc.gov/standards/mods/) for individual digital object descriptions and OAI-ORE Resource Maps for the creation of aggregated objects.

[636] CRUD-The creation, update or deletion of an object in a digital repository.

[637] The discussion over how to both design and implement persistent and unique identifiers has a vast body of literature. For some recent work, see (Tonkin 2008), (Campbell 2007), and (Hilse and Kothe 2006).

architecture will be based on the "Archaeological Records of Europe - Networked Access" (ARENA)." [638] The second will be a Fedora demonstrator [639] that will "integrate the access, archiving, harvesting and organization of electronic resources." As Blanke has explained, "the central aim of the demonstrator will be the evaluation of DARIAH as a robust and flexible infrastructure that allows easy exchange of its components"(Blanke 2010). Similar to TextGrid, DARIAH intends to create a flexible architecture that is loosely coupled so other communities can also add their own services on top of it.

Another important factor considered by DARIAH is the frequently distributed nature of humanities data, for example, one digital archive may have transcriptions of a manuscript while another digital archive has digital images of this manuscript. Thus DARIAH plans to build a data architecture that will "cover the easy exchange of file type data, the ability to create relationships between files in remote locations and flexible caching mechanism to deal with the exchange of large single data items like digitization images"(Blanke 2010). Since humanities data also needs to be preserved for long periods of time in order to support both archiving and reuse, DARIAH also plans to incorporate existing archived research data. In his final overview of the project, Blanke proposed that:

> DARIAH is one way to build a research infrastructure for the humanities. It uses grid technologies together with digital library technologies to deliver services to support the information needs of humanities researchers. It integrates many services useful for humanities research and will focus less on automation of processing but on providing an infrastructure to support the main activity of humanities researchers, the attempt to establish the meaning of textual and other human created resources (Blanke 2010).

He also reported that DARIAH is being built on existing national infrastructures and will consequently be "embedded" among service providers that are already in place in order to ensure the best possible chances of success.

Digital Humanities Observatory

The Digital Humanities Observatory (DHO) [640] is a digital humanities "collaboratory working with Humanities Serving Irish Society (HSIS), national, European, and international partners to further e-scholarship." This organization was founded in 2008 and was created to support digital humanities research in Ireland and help manage the creation and preservation of digital resources. The DHO will focus on three main issues in the next few years: "encouraging collaboration; providing for the management, access, and preservation of project data; and promulgating shared standards and technology for project development"(Schreibman et al. 2009). The DHO has also pointed out that the expectations of digital humanities centers are rapidly changing and plans for long-term viability need to be created from the very *beginning* of development.

Thus Schreibman et al. also noted that the creation of the DHO is in line with other growing initiatives such as Project Bamboo and DARIAH, where digital humanities projects are moving away from "digital silos" to an approach where scholarly resources will be "linked, interoperable, reusable, and preserved." The DHO is currently creating three "distinct but integrated infrastructures": 1) a portal that will serve as the public face of the DHO and is based on the Drupal [641] content management system; 2) DRAPIER (Digital Research and Practices in Ireland), which is also based on Drupal, and will serve as a framework for public discovery of digital projects in Ireland; and 3) an "access and preservation repository based on Fedora." Some resources created by HSIS partners of the DHO will reside in their Fedora repository while others will be federated in Fedora instances managed by DHO partners.

---

[638] http://www.dariah.eu/index.php?option=com_content&view=article&id=30&Itemid=34

[639] From the website (May 13, 2010), it appears that this Fedora demonstrator may have been reconceptualized as the TEI Demonstrator, "The purpose of the DARIAH TEI Demonstrator is to demonstrate the practical benefits of using TEI for the representation of digital resources of all kinds, but primarily of original source collections within the arts and humanities. The Demonstrator aims to make it easy for humanities researchers to share TEI-encoded texts with others, and to compare their encoding practice with that of others in the TEI community." This demonstrator will use a "special purpose software platform" developed by the Max Planck Digital Library called eSciDoc (https://www.escidoc.org/) that as part of its core functionality includes a Fedora repository.

[640] http://dho.ie/

[641] http://drupal.org/

DRIVER

DRIVER (Digital Repository Infrastructure Vision for European Research) [642] is a "multi-phase effort whose vision and primary objective is to establish a cohesive, pan-European infrastructure of digital repositories, offering sophisticated functionality services to both researchers and the general public." At the end of its first stage in November 2007, the DRIVER project provided access to a testbed system that produced a search portal with open access content from over 70 repositories. The DRIVER efforts initially concentrated on the infrastructure aspect and developed "clearly defined interfaces to the content network, which allow any qualified service-provider to build services on top of it." [643] As of March 2010, the DRIVER search portal offered access to over 2,500,000 publications from 249 repositories in 39 countries. In its current stage DRIVER II seeks to expand its geographic coverage, support more advanced end-user functionality for searching complex digital objects and provide access to a greater variety of open access materials.

NoC-Network of Expert Centres

The Network of Expert Centres [644] is a collaboration of "centres with expertise in digital arts and humanities research and scholarship, including practice-led research." The areas of research expertise include "data creation, curation, preservation, management (including rights and legal issues), access and dissemination, and methodologies of data use and re-use." Membership in NoC is open to centres in Great Britain and Ireland that have a formal institutional status, recognized expertise in digital arts and humanities, a history of persistence, and an institutional or inter-institutional focus of activity. The current list of participating organizations includes the Archaeology Data Service (ADS), the Centre for Computing in the Humanities (CCH), the Centre for Data Digitization and Analysis (CDDA), the Digital Design Studio (DDS), the History Data Service (HDS), the Humanities Advanced Technology & Information Institute (HATII), the Humanities Research Institute (HRI), the Oxford Text Archive (OTA), the UCL Centre for Digital Humanities, and the VADS (Visual Arts Data Service).

The purpose of this network is to enable all the members to pursue a series of collective aims and objectives in the support of arts and humanities research and scholarship, for much of which arts-humanities.net will provide a central hub. These objectives include: 1) promoting the broad use of ICT; 2) providing leadership in the use of digital methods and resources; 3) developing and exchanging expertise, standards and best practices; 4) identifying and serving the needs of the research community; and 5) conducting dialogue with stakeholders. NoC has a steering committee of six voting members drawn from the different centres and their role is to propose initiatives and activities, to provide reports to the membership, to convene regular meetings of the full network, and to convene workgroups.

Project Bamboo

According to its website Project Bamboo [645] "is a multi-institutional, interdisciplinary, and inter-organizational effort that brings together researchers in arts and humanities, computer scientists, information scientists, librarians, and campus information technologists" in order to answer one major question: "How can we advance arts and humanities research through the development of shared technology services?" Project Bamboo was designed as a "community driven cyberinfrastructure initiative" (Kainz 2009) and received funding from the Mellon Foundation as a planning project and it will conclude in September 2010. While Project Bamboo's main task is to defined shared technology services they also hope to identify "organizational, partnership and social" models.

The Bamboo project plans to eventually submit an implementation proposal to Mellon and hopes to initiate a one year project as part of a 10 year program. In this first year Bamboo proposes to focus on developing scholarly networking services, the Bamboo atlas and the Bamboo services platform. In terms of scholarly

[642] http://www.driver-repository.eu/
[643] http://www.driver-repository.eu/Driver-About/About-DRIVER.html
[644] http://www.arts-humanities.net/noc
[645] http://projectbamboo.org/

networking, Bamboo plans to develop "gadgets" or small components that plug into existing VREs and social platforms. Secondly, the Bamboo atlas will define a "collection of services." Finally, the Bamboo services platform will "establish a foundation technology to host and deliver shared services for arts and humanities research, teaching and learning" where "services shall run on geographically distributed yet interlinked instances of the platform ("cloud") that shall enable "always-available" guarantees to service adopters" (Kainz 2009). Bamboo plans to adopt common standards and *reuse* rather than develop its own services and technology. Initial demonstrators will be created to test a number of their findings.

The Bamboo Project held five workshops between the fall of 2008 and the summer of 2009, with over 600 individuals participating from a number of organizations and institutions. [646] Workshop One entitled "*The Planning Process & Understanding Arts and Humanities Scholarship*" involved four individual workshops at different locations where scholars in the arts and humanities held dialogues with information technologists to understand the scholarship practices of these disciplines and their future directions. One major goal of these workshops was to develop a "high-level list of scholarly practices related to arts and humanities" and this list as well as all individual workshop notes, documents and other materials were placed online in the project planning wiki. [647]

Workshop Two built off the results of the first series of workshops and then examined possible future directions for Project Bamboo including: advocacy and leadership, education and training, institutional partnerships and support, scholarly networking, standards and best practices, "tools, repositories and content partners", and a shared service framework. At the end of this workshop, seven working groups [648] were formed around these themes, with the addition of an eighth working group entitled "Stories" which was chartered to "collect narratives and/or illustrative examples on behalf of all Bamboo working groups that express particular aspects of scholarship, scholarly workflow, research, and/or teaching that are or could be facilitated by technology." Since the Bamboo project, however, is still in the implementation phase, the "Principles for Leadership," and "Standards and Best Practices," workgroups have yet to begin their tasks. Each active working group has a separate wiki page that allows group members to collaborate, post documents, and provide other information. Working groups were also charged with creating demonstrators [649] "to support the discussion and analysis activities of the working groups, and to reflect and test the results of that discussion and analysis." [650] Initial progress of the working groups was reported at Workshop Three, and a "straw consortial model" for the project was also introduced. The fourth workshop involved the discussion of a draft program document and the fifth workshop finalized plans for an implementation proposal to Mellon.

The amount of data generated from these workshops and found on the project wiki is quite extensive and the Bamboo Project has recently begun to synthesize this data [651] and make it available on the website, organized both by the major topics and by workshop. A "Bamboo Analysis" [652] section of the website has also been created to present the initial results of the data analysis. This analysis has produced a list of major themes (contextualization, annotation, tenure, credit and peer review, "how research and technology overlap with pedagogy") and each of these themes includes a separate analysis page with relevant quotes drawn from workshop participants.

SEASR

SEASR, or the Software Environment for the Advancement of Scholarly Research (SEASR), has been funded by the Mellon Foundation as a "transformational cyberinfrastructure technology" and seeks to support two major functions: 1) to enable scholars to both individually and collaboratively pursue computationally advanced

[646] A sixth workshop was held in June of 2010.
[647] https://wiki.projectbamboo.org/
[648] http://projectbamboo.org/working-groups-ws2-ws3
[649] A list of demonstrators can be found on the Project Planning Wiki (https://wiki.projectbamboo.org/display/BPUB/Demonstrators+List)
[650] https://wiki.projectbamboo.org/display/BPUB/About+Demonstrators
[651] http://projectbamboo.org/planningproject/data
[652] http://projectbamboo.org/planningproject/analysis

digital research in a robust virtual work environment; 2) to support digital humanities developers with a robust programming environment where they can both rapidly and efficiently design applications that can be shared.

To begin with, SEASR provides a visual programming environment named Meandre [653] that allows its users to develop applications that are labeled "flows" that can then be deployed on an already existing robust hardware infrastructure. According to the project website Meandre is a "semantic enabled web-driven, dataflow execution environment" that both provides "machinery for assembling and executing data flows -software applications consisting of software components that process data" as well as "publishing capabilities for flows and components, enabling users to assemble a repository of components for reuse and sharing." In other words, digital humanities developers can use Meandre to both quickly develop and share software applications to support both individual scholarship and research collaboration as well as reuse applications that have been developed by others, as SEASR maintains an expanding repository of different components and applications.

The second major function of SEASR is to provide a virtual work environment where digital humanities scholars can share both data and research and provides a variety of data and text mining tools including frequent pattern mining, clustering, text summarization, information extraction, and named entity recognition. This work environment allows scholars to access digital materials that are stored in a variety of formats, experiment with a variety of different algorithms, and utilize supercomputing power to provide new visualizations and discover new relationships between data.

SEASR uses both a service oriented architecture (SOA) and semantic web computing [654] to address four key research needs: 1) the ability to transform semi or unstructured data (including natural language texts) into structured data; 2) to improve automatic knowledge discovery through analytics; 3) to support collaborative scholarship through a virtual research environment; and 4) to promote open source development and community involvement through sharing user applications developed through Meandre in a community repository.

A number of digital humanities projects have utilized SEASR in their work including the Networked Environment for Music Analysis (NEMA) [655] and the MONK (Metadata Offer New Knowledge) project [656]

TextGrid

TextGrid first began work in 2006 and has now evolved into a joint project of ten partners with funding through 2012. The project is working to create an infrastructure for a VRE in the humanities that consists of two key components: 1) the TextGrid repository that will serve as a "long-term archive for research data in the humanities, embedded in a grid infrastructure" and will "ensure long-term availability and access to its research data as well as interoperability" and 2) the "TextGrid Laboratory" that will serve as the point of entry to the VRE and provide access to both existing and new tools. [657] TextGrid is soliciting continuous feedback on the TextGrid laboratory in order to improve it, add new tools and address interface issues. A beta version of TextGrid laboratory can also currently be downloaded. [658]

The TextGrid project has published extensively on their work and that literature will briefly be reviewed here. TextGrid's initial audience was philologists and their early work established a "community grid for the collaborative editing, annotation, analysis and publication of specialist texts" (Aschenbrenner et al. 2009). Initial research conducted by TextGrid largely focused on this aspect (Dimitriadis et al. 2006, Gietz et al. 2006) of the project, particularly in the development of philological editions and services for scholars using them. More recent research by the TextGrid group has presented detailed information on the technical architecture of the project and how it relates to the larger world of eHumanities and cyberinfrastructure (Aschenbrenner et al. 2009, Ludwig and Küster 2008, Zielinski et al. 2009).

[653] http://seasr.org/meandre/documentation/
[654] http://seasr.org/documentation/overview/
[655] http://www.music-ir.org/?q=node/12
[656] http://monkproject.org/. For more on their use of SEASR in text mining and named entity recognition see (Vuillemot et al. 2009).
[657] http://www.textgrid.de/en/ueber-textgrid.html
[658] http://www.textgrid.de/en/beta.html

The creators of TextGrid maintain that in the humanities key resources are data, "knowledge about data" and services where the consequent challenge is to process and connect these resources (Aschenbrenner et al. 2009). The intellectual content and the community that uses it are at the core of TextGrid. While the majority of content within TextGrid will be textual, image resources have also been provided by a number of German institutions and these two resources will then be merged into a "virtual library" using the Globus toolkit grid infrastructure. This will allow TextGrid to provide seamless searching over federated archives while still allowing the data to remain *distributed* and also support the addition of new organizations (Ludwig and Küster 2008). The authors also discussed the difficulties of creating digital content that will need to be used and preserved for a time that will *outlast* any system design:

> Furthermore, the typical project duration of eHumanities projects and in particular that for the elaboration of critical editions, academic dictionaries and large linguistic corpora is often long — many years at least, often decades, sometimes even centuries. The time-span during which those resources remain pertinent for then current research can be much longer still. In this it by far surpasses the average lifetime of any particular generation of software technology (Ludwig and Küster 2008).

They thus pointed out the importance of content stability and also of designing content that can be ported into new systems as time passes, since digital resources created will likely be used for far longer than the individual services designed to use them.

TextGrid's developers recommended creating an open service environment with robust and general services, which can ultimately be used to "form the basis for value added services and, eventually, domain specific services and tailored applications" (Aschenbrenner et al. 2009). The creators of TextGrid thus argued that in order to be successful a digital humanities project must first create an *open* infrastructure with fairly *generic* services to begin while at the same time promoting *community* creation of specialized applications and workflows that can motivate community participation and greater uptake. In fact, active community building has been one of TextGrid's most dynamic tasks during both project design and development, and they have designed specific documentation and communication for its three intended users groups (visitors/users, data providers and tool developers).

TextGrid has designed a service oriented architecture in multiple layers: 1) the application environment or TextGrid access point that is Eclipse based and geared towards philologists; 2) services, or "building blocks of specialized functionality" including functionalities such as tokenization, lemmatization and collation, that are "wrapped into individual services to be re-used by other services or plugged into an application environment;" 3) the TextGrid middleware; and 4) stable archives (Aschenbrenner et al. 2009). They have also developed a semantic service registry for TextGrid. Zielinski et al. have offered a concise summary of their approach:

> The TextGrid infrastructure is a multilayered system created with the motivation to hide the complex grid infrastructure…from the scholars and to make it possible to integrate external services with TextGrid tools. Basically in this service oriented architecture (SOA), there are three layers: the user interface, a services layer with tools for textual analysis and text processing, and the TextGrid middleware, which itself includes multiple layers (Zielinksi et al. 2009).

Nonetheless the TextGrid project faced a variety of data interoperability challenges within their own project in terms of using the TEI as its basic form of markup, since a variety of partner projects used the TEI at varying levels of sophistication. While they did not want to sacrifice the depth of semantic encoding the TEI offered, they at the same time needed to define a minimum "abstraction level" necessary to promote larger interoperability of computational processes in TextGrid (Blanke et al. 2008). As a solution, TextGrid developed a "core" encoding approach:

> … which follows a simple principle: one can always go from a higher semantic degree to a lower semantic degree; and in possession of a suitable transformation script, this mapping can be done automatically. TextGrid encourages all its participating projects to describe their data in an XML-based markup that is suitable for their specific research questions. At the same time projects can register a mapping from their specific, semantically deep data to the respective TextGrid-wide core encoding that is a reasonably expressive TEI-subset (Blanke et al. 2008)

TextGrid's solution thus attempts to respect the sophisticated encoding of local practices while still maintaining a basic level of interoperability. This illustrates the difficulties of supporting cross corpora searching even within *one* project. All content that is created either with the help of TextGridLab or comes from external

resources is saved unchanged to the TextGrid repository, metadata is extracted and normalized before being stored in central metadata storage and a full text index is also extracted from the raw data repository and updated with all changes (Ludwig and Küster 2008).

The TextGridLab tool (an Eclipse based GUI) is intended to help users create TEI resources that can live within the data grid. Although TEI documents form a large part of the resources in TextGrid it can handle heterogeneous data formats (plain text, TEI/XML, images). TextGrid also provides a number of basic services (tokenizers, lemmatizers, sorting tools, streaming editors, collation tools) that can be used against its objects while also letting users create their own services within a Web services framework:

> Web Service frameworks are available for many programming languages—so if a person or institution wishes to make his/her text processing tool available to the TextGrid community and the workflow engine, the first step is to implement a Web Service wrapper for the tool and deploy it on a public server (or one of TextGrid's). The next steps are to apply for registration in the TextGrid service registry and to provide a client plug-in for the Eclipse GUI so that the tool is accessible for humans (GUI) and machines (service registry) alike (Zielinski et al. 2009).

This architecture makes it easy to extend the TextGrid framework to work with other digital humanities applications. Further details on how users can search the TextGrid collections can be found in (Zielinski et al. 2009)

TextVRE

TextVRE [659] is a recently initiated project by the Center for e-Research (CeReh) and the Center for Computing in the Humanities at King's College London, the University of Sheffield Natural Language Processing Group, and the State and University Library at Göttingen. According to their project website:

> The overall aim of TEXTvre is to support the complete lifecycle of research in e-Humanities textual studies. The project provides researchers with advanced services to process and analyse research texts that are held in formally managed, metadata-rich institutional repositories. The access and analysis of textual research data will be supported by annotation and retrieval technology and will provide services for every step in the digital research life cycle.

The TEXTvre will build off of the results of the TextGrid project, but will be adapted to UK needs and bring together the major organizations in "e-Humanities textual studies." The project plans to embed itself within the daily workflow of researchers at King's College and to be interoperable with institutional repository and data management structures. This project is currently at its very beginning stages, but does have a useful list of tools for potential inclusion in the "virtual research environment" they will develop. [660]

References

[ACLS 2006]. American Council of Learned Societies. *Our Cultural Commonwealth: The Final Report of the American Council Of Learned Societies Commission on Cyberinfrastructure for the Humanities & Social Sciences*. ACLS, (2006). http://www.acls.org/cyberinfrastructure/

[Agosti et al. 2005]. Agosti, Maristella, Nicola Ferro, and Nicola Orio. "Annotating Illuminated Manuscripts: an Effective Tool for Research and Education." *JCDL '05: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY, USA, ACM (2005): 121-130. http://dx.doi.org/10.1145/1065385.1065412

[Amin et al. 2008]. Amin, Alia, Jacco van Ossenbruggen, Lynda Hardman, and Annelies van Nispen. "Understanding Cultural Heritage Experts' Information Seeking Needs." *JCDL '08: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY, USA: ACM, (2008): 39-47. http://dx.doi.org/10.1145/1378889.1378897

[659] http://textvre.cerch.kcl.ac.uk/
[660] http://textvre.cerch.kcl.ac.uk/?page_id=9

[APA/AIA 2007]. *APA/AIA Task Force on Electronic Publications: Final Report*. American Philological Institution/American Archaeological Institute of America, Philadelphia, PA; Boston, MA, (March 2007). http://socrates.berkeley.edu/~pinax/pdfs/TaskForceFinalReport.pdf

[ARL 2009a]. Association of Research Libraries. "Establish a Universal, Open Library or Digital Data Commons." Association of Research Libraries, (January 2009). http://www.arl.org/bm~doc/ibopenlibpsa2.pdf

[ARL 2009b]. Association of Research Libraries. *The Research Library's Role in Digital Repository Services: Final Report of the ARL Digital Repository Issues Task Force*. Association of Research Libraries, (January 2009). http://www.arl.org/bm~doc/repository-services-report.pdf

[Arms and Larsen 2007]. Arms, William Y. and Ronald L. Larsen. *The Future of Scholarly Communication: Building the Infrastructure for Cyberscholarship*. National Science Foundation, (September 2007). http://www.sis.pitt.edu/~repwkshop/SIS-NSFReport2.pdf

[Aschenbrenner et al. 2008]. Aschenbrenner, Andreas, Tobias Blanke, David Flanders, Mark Hedges, and Ben O'Steen. "The Future of Repositories? Patterns for (Cross-)Repository Architectures." *D-Lib Magazine*, 14 (November 2008). http://www.dlib.org/dlib/november08/aschenbrenner/11aschenbrenner.html

[Aschenbrenner et al. 2009]. Aschenbrenner, Andreas, Marc W. Küster, Christoph Ludwig and Thorsten Vitt. "Open eHumanities Digital Ecosystems and the Role of Resource Registries." *DEST '09: 3rd IEEE International Conference Digital Ecosystems and Technologies*, (June 2009): 745-750. http://dx.doi.org/10.1109/DEST.2009.5276672

[Aschenbrenner et al. 2010]. Aschenbrenner, Andreas, Tobias Blanke, and Marc W. Küster. "Towards an Open Repository Environment." *Journal of Digital Information*, 11 (March 2010). http://journals.tdl.org/jodi/article/view/758

[Ashdowne 2009]. Ashdowne, Richard. "Accidence and Acronyms: Deploying Electronic Assessment in Support of Classical Language Teaching in a University Context." *Arts and Humanities in Higher Education*, 8 (June 2009): 201-216.

[Audenaert and Furuta 2010]. Audenaert, Neal and Richard Furuta. "What Humanists Want: How Scholars Use Source Materials." *JCDL '10: Proceedings of the 10th Annual Joint Conference on Digital Libraries*. New York, NY, USA: ACM, (2010): 283-292. http://dx.doi.org/10.1145/1816123.1816166

[Bagnall 2010]. Bagnall, Roger. "Integrating Digital Papyrology." Paper presented at *Online Humanities Scholarship: The Shape of Things to Come*, (March 2010). http://hdl.handle.net/2451/29592

[Baker et al. 2008]. Baker, Mark, Claire Fisher, Emma O'Riordan, Matthew Grove, Michael Fulford, Claire Warwick, Melissa Terras, Amanda Clarke, and Mike Rains. "VERA: A Virtual Environment for Research in Archaeology." *Fourth International Conference on e-Social Science*, University of Manchester, (June 2008). http://www.ncess.net/events/conference/programme/fri/3abaker.pdf

[Ball 2010]. Ball, Alex. *Preservation and Curation in Institutional Repositories (Version 1.3)*. Digital Curation Centre, UKOLN, University of Bath, (March 2010). http://dcc.ac.uk/sites/default/files/documents/reports/irpc-report-v1.3.pdf

[Bamman and Crane 2006]. Bamman, David and Gregory Crane. "The Design and Use of a Latin Dependency Treebank." *TLT 2006: Proceedings of the Fifth International Treebanks and Linguistic Theories Conference*, (2006): 67-78. http://hdl.handle.net/10427/42684

[Bamman and Crane 2007]. Bamman, David and Gregory Crane. "The Latin Dependency Treebank in a Cultural Heritage Digital Library." *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTech 2007)*, (2007): 33-40. http://acl.ldc.upenn.edu/W/W07/W07-0905.pdf

[Bamman and Crane 2008]. Bamman, David and Gregory Crane. "Building a Dynamic Lexicon from a Digital Library." *JCDL '08: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY, USA: ACM, (2008): 11-20. Preprint available at http://hdl.handle.net/10427/42686

[Bamman and Crane 2009]. Bamman, David and Gregory Crane. "Computational Linguistics and Classical Lexicography." *Digital Humanities Quarterly*, 3 (January 2009). http://www.digitalhumanities.org/dhq/vol/3/1/000033.html

[Bamman, Mambrini and Crane 2009]. Bamman, David, Francesco Mambrini, and Gregory Crane. "An Ownership Model of Annotation: The Ancient Greek Dependency Treebank." *TLT 2009-Eighth International Workshop on Treebanks and Linguistic Theories*, Milan, Italy, (November 2009). Preprint available at: http://www.perseus.tufts.edu/publications/tlt8.pdf

[Bamman, Passarotti, and Crane 2008]. "A Case Study in Treebank Collaboration and Comparison: *Accusativus cum Infinitivo* and Subordination in Latin." *Prague Bulletin of Mathematical Linguistics*, 90 (December 2008): 109-122. http://ufal.mff.cuni.cz/pbml/90/art-bamman-et-al.pdf

[Bankier and Perciali 2008]. Bankier, Jean-Gabriel and Irene Perciali. "The Institutional Repository Rediscovered: What Can a University Do for Open Access Publishing." *Serials Review*, 34 (March 2008): 21-26. Also available at: http://works.bepress.com/jean_gabriel_bankier/1

[Barker 2010]. Barker, Elton. "Repurposing *Perseus*: the Herodotus Encoded Space-Text-Image Archive (HESTIA)." Presentation given at NEH-DFG Workshop, Medford, MA, (January 2010). http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/events-en/nehdfg/pdf/hestia-presentation

[Barker et al. 2010]. Barker, Elton, Stefan Bouzarovski, Chris Pelling, and Leif Isaksen. "Mapping an Ancient Historian in a Digital Age: the Herodotus Encoded Space-Text-Image Archive (HESTIA)." *Leeds International Classical Studies*, 9 (March 2010). http://www.leeds.ac.uk/classics/lics/2010/201001.pdf

[Barmpoutis et al. 2009]. Barmpoutis, Angelos, Eleni Bozia, and Robert S. Wagman. "A Novel Framework for 3D Reconstruction and Analysis of Ancient Inscriptions." *Machine Vision and Applications*, 21 (2009): 989-998. http://dx.doi.org/10.1007/s00138-009-0198-7

[Barrenechea 2006]. Barrenechea, Francisco. "A Fragment of Old Comedy: P. Columbia inv. 430." *Zeitschrift für Papyrologie und Epigraphik*, 158 (2006): 49-54. http://www.jstor.org/stable/20191149

[Bauer et al. 2008]. Bauer, Péter, Zsolt Hernáth, Zoltán Horváth, Gyula Mayer, Zsolt Parragi, Zoltán Porkoláb, and Zsolt Sztupák. "HypereiDoc – An XML Based Framework Supporting Cooperative Text Editions." *Advances in Databases and Information Systems* (*Lecture Notes in Computer Science*, Volume 5207), (2008): 14-29. http://dx.doi.org/10.1007/978-3-540-85713-6_3

[Baumann and Seales 2009]. Baumann, Ryan and Brent W. Seales. "Robust Registration of Manuscript Images." *JCDL '09: Proceedings of the 2009 Joint International Conference on Digital Libraries*. New York, NY, USA: ACM, (2009): 263-266. http://dx.doi.org/10.1145/1555400.1555443

[Beacham and Denard 2003]. Beacham, Richard and Hugh Denard. "The Pompey Project: Digital Research and Virtual Reconstruction of Rome's First Theatre." *Computers and the Humanities*, 37 (February 2003): 129-139. http://dx.doi.org/10.1023/A:1021859830043

[Benardou et al. 2010a]. Benardou, Agiatis, Panos Constantopoulos, Costis Dallas, and Dimitris Gavrilis. "A Conceptual Model for Scholarly Research Activity." *iConference 2010*, (February 2010). http://hdl.handle.net/2142/14945

[Benardou et al. 2010b]. Benardou, Agiatis, Panos Constantopoulos, Costis Dallas, and Dimitris Gavrilis. "Understanding the Information Requirements of Arts and Humanities Scholarship." *International Journal of Digital Curation*, 5 (July 2010). http://ijdc.net/index.php/ijdc/article/view/144/0

[Berti et al. 2009]. Berti, Monica, Matteo Romanello, Alison Babeu, and Gregory Crane. "Collecting Fragmentary Authors in a Digital Library." *JCDL '09: Proceedings of the 2009 Joint International Conference on Digital Libraries*. New York, NY, USA: ACM, (2009): 259-262. Preprint available at: http://www.perseus.tufts.edu/publications/JCDL09_sp.pdf

[Bilane et al. 2008]. Bilane, Petra, Stéphane Bres, and Hubert Emptoz. "Local Orientation Extraction for Wordspotting in Syriac Manuscripts." *Image and Signal Processing* (*Lecture Notes in Computer Science*, Volume 5099), (2008): 481-489. http://dx.doi.org/10.1007/978-3-540-69905-7_55

[Bizer et al. 2007]. Bizer, Chris, Richard Cyganiak, and Tom Heath. "How to Publish Linked Data on the Web?" http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial/

[Blackwell and Crane 2009]. Blackwell, Christopher and Gregory Crane. "Conclusion: Cyberinfrastructure, the Scaife Digital Library and Classics in a Digital Age." *Digital Humanities Quarterly*, 3 (January 2009), http://www.digitalhumanities.org/dhq/vol/3/1/000035.html

[Blackwell and Martin 2009]. Blackwell, Chris and Thomas R. Martin. "Technology, Collaboration, and Undergraduate Research." *Digital Humanities Quarterly*, 3 (January 2009). http://www.digitalhumanities.org/dhq/vol/3/1/000024.html

[Blackwell and Smith 2009]. Blackwell, Christopher and David Neel Smith. "Homer Multitext - Nine Year Update." *Digital Humanities 2009 Conference Abstracts*, (June 2009): 6-8. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf

[Blanke 2010]. Blanke, Tobias. "From Tools and Services to e-Infrastructure for the Arts and Humanities." *Production Grids in Asia*, Part II (2010): 117-127. http://dx.doi.org/10.1007/978-1-4419-0046-3_10

[Blanke et al. 2008]. Blanke, Tobias, Andreas Aschenbrenner, Marc Küster, and C. Ludwig. "No Claims for Universal Solutions - Possible Lessons from Current e-Humanities Practices in Germany and the UK." *E-SCIENCE '08: Proceedings of the IEEE e-Humanities Workshop*, (November 2008). http://www.clarin.eu/system/files/ClaimsUniversal-eHum2008.pdf

[Blanke et al. 2009]. Blanke, Tobias, Mark Hedges, and Stuart Dunn. "Arts and Humanities E-Science—Current Practices and Future Challenges." *Future Generation Computer Systems*, 25 (April 2009): 474-480. http://dx.doi.org/10.1016/j.future.2008.10.004

[Blanke, Hedges and Palmer 2009]. Blanke, Tobias, Mark Hedges, and Richard Palmer. "Restful Services for the e-Humanities — Web Services that Work for the e-Humanities Ecosystem." *3rd IEEE International Conference on Digital Ecosystems Technologies (DEST 2009)*, (June 2009): 637-642.

http://dx.doi.org/10.1109/DEST.2009.5276740

[Bodard 2006]. Bodard, Gabriel. "Inscriptions of Aphrodisias: Paradigm of an Electronic Publication." *CLiP 2006: Literatures, Languages and Cultural Heritage in a Digital World*, (July 2006). http://www.cch.kcl.ac.uk/clip2006/redist/abstracts_pdf/paper33.pdf

[Bodard 2008]. Bodard, Gabriel. "The *Inscriptions of Aphrodisias* as Electronic Publication: A User's Perspective and a Proposed Paradigm." *Digital Medievalist*, 4 (2008). http://www.digitalmedievalist.org/journal/4/bodard/

[Bodard 2009]. Bodard, Gabriel. "Digital Classicist: Re-use of Open Source and Open Access Publications in Ancient Studies." *Digital Humanities 2009 Conference Abstracts*, (June 2009): 2. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf

[Bodard and Garcés 2009]. Bodard, Gabriel and Juan Garcés. "Open Source Critical Editions: A Rationale." in *Text Editing, Print and the Digital World.* Ashgate Publishing, 2009, pp. 83-98.

[Bodard and Mahony 2010]. Bodard, Gabriel and Simon Mahony, eds. *Digital Research in the Study of Classical Antiquity*. (Digital Research in the Arts and Humanities Series). Burlington, VT: Ashgate Publishing, 2010.

[Bodard et al. 2009]. Bodard, Gabriel, Tobias Blanke, and Mark Hedges. "Linking and Querying Ancient Texts: a Case Study with Three Epigraphic/Papyrological Datasets." *Digital Humanities 2009 Conference Abstracts*, (June 2009): 2-4. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf

[Bolter 1991]. Bolter, Jay D. "The Computer, Hypertext, and Classical Studies." *The American Journal of Philology*, 112 (1991): 541-545. http://dx.doi.org/10.2307/294933

[Borgman 2009]. Borgman, Christine L. "The Digital Future is Now: A Call to Action for the Humanities." (2009), http://works.bepress.com/cgi/viewcontent.cgi?article=1232&context=borgman.

[Boschetti 2007]. Boschetti, Federico. "Methods to Extend Greek and Latin Corpora with Variants and Conjectures: Mapping Critical Apparatuses onto Reference Text." *CL 2007: Proceedings of the Corpus Linguistics Conference*, University of Birmingham, UK, (July 2007). http://corpus.bham.ac.uk/corplingproceedings07/paper/150_Paper.pdf

[Boschetti 2009]. Boschetti, Federico. "Digital Aeschylus - Breadth and Depth Issues in Digital Libraries." *Workshop on Advanced Technologies for Digital Libraries 2009 (AT4DL 2009)*, (September 2009): 5-8. http://www.unibz.it/en/public/universitypress/publications/all/Documents/9788860460301.pdf

[Boschetti et al. 2009]. Boschetti, Federico, Matteo Romanello, Alison Babeu, David Bamman, and Gregory Crane. "Improving OCR Accuracy for Classical Critical Editions." *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, Corfu, Greece, (2009): 156-167. Preprint available at: http://www.perseus.tufts.edu/publications/ecdl2009-preprint.pdf

[Bowman et al. 2009]. Bowman, Alan K., R. S. O. Tomlin and Klaas Worp. "Emptio Bovis Frisica: The 'Frisian Ox Sale' Reconsidered." *Journal of Roman Studies*, 99 (2009): 156-70.

[Bowman et al. 2010]. Bowman, Alan K., Charles V. Crowther, Ruth Kirkham, and John Pybus. "A Virtual Research Environment for the Study of Documents and Manuscripts." In *Digital Research in the Study of Classical Antiquity* (eds. Gabriel Bodard and Simon Mahony). Ashgate Publishing, 2010, pp. 87-103.

[Bozzi and Calabretto 1997]. Bozzi, Andrea and Sylvie Calabretto. "The Digital Library and Computational Philology: The BAMBI Project." *Research and Advanced Technology for Digital Libraries*, (*Lecture Notes in Computer Science*, Volume 1324), (1997): 269-285. http://dx.doi.org/10.1007/BFb0026733

[Bradley 2005]. Bradley, John. "What You (Fore)see is What You Get: Thinking about Usage Paradigms for Computer Assisted Text Analysis." *Text Technology*, 14 (2005): 1-19. Also available online: http://texttechnology.mcmaster.ca/pdf/vol14_2/bradley14-2.pdf

[Bradley 2008]. Bradley, John. "*Pliny*: A Model for Digital Support of Scholarship." *Journal of Digital Information*, 9 (May 2008). http://journals.tdl.org/jodi/article/view/209/198

[Bradley and Short 2005]. Bradley, John and Harold Short. "Texts into Databases: The Evolving Field of New-style Prosopography." *Literary & Linguistic Computing*, 20 (January 2005): 3-24. http://dx.doi.org/10.1093/llc/fqi022

[Breuel 2009]. Breuel, Thomas. "Applying the OCRopus OCR System to Scholarly Sanskrit Literature." *Sanskrit Computational Linguistics (Lecture Notes in Computer Science*, Volume 5402), (2009): 391-402. http://dx.doi.org/10.1007/978-3-642-00155-0_21

[Brown and Greengrass 2010]. Brown, Stephen and Mark Greengrass. "Research Portals in the Arts and Humanities." *Literary & Linguistic Computing*, 25 (April 2010): 1-21. http://dx.doi.org/10.1093/llc/fqp032

[Brunner 1991]. Brunner, Theodore F. "The Thesaurus Linguae Graecae: Classics and the Computer." *Library Hi-Tech*, 9 (1991): 61-67.

[Buchanan et al. 2005]. Buchanan, George, Sally Jo Cunningham, Ann Blandford, Jon Rimmer, and Claire Warwick. "Information Seeking by Humanities Scholars." *ECDL '05: Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries*. Springer, (2005): 218-229.

[Buchanan 2010]. Buchanan, Sarah. "Accessioning the Digital Humanities: Report from the 1st Archival Education and Research Institute." *Digital Humanities Quarterly*, 4 (August 2010), http://www.digitalhumanities.org/dhq/vol/4/1/000084/000084.html

[Büchler et al. 2008]. Büchler, Marco, Gerhard Heyer, and Sabine Grunder. "eAQUA - Bringing Modern Text Mining Approaches to Two Thousand Years Old Ancient Texts." *e-Humanities – an emerging discipline: Workshop in the 4th IEEE International Conference on e-Science*, (December 2008). http://www.clarin.eu/system/files/2008-09-05-IEEE2008-eAQUA-project.pdf

[Büchler and Geßner 2009]. Büchler, Marco and Annette Geßner. "Citation Detection and Textual Reuse on Ancient Greek Texts." *DHCS 2009-Chicago Colloquium on Digital Humanities and Computer Science*, Chicago, (November 2009). http://lingcog.iit.edu/%7Eargamon/DHCS09-Abstracts/Buechler-Gessner.pdf

[Bulacu and Schomaker 2007]. Bulacu, Marius and Lambart Schomaker. "Automatic Handwriting Identification on Medieval Documents." *ICIAP 2007: 14th International Conference on Image Analysis and Processing*, (2007): 279-284. http://www.ai.rug.nl/~bulacu/iciap2007-bulacu-schomaker.pdf

[Bulst 1989]. Bulst, Neithard. "Prosopography and the Computer: Problems and Possibilities." in *History and Computing II* (ed. Peter Denley). Manchester, UK: Manchester University Press, 1989, pp. 12–18. Postprint available at http://repositories.ub.uni-bielefeld.de/biprints/volltexte/2010/4053

[Calanducci et al. 2009]. Calanducci, Antonio, Jorge Sevilla, Roberto Barbera, Giuseppe Andronico, Monica Saso, Alessandro De Filippo, Stefania Iannizzotto, Domenico Vicinanza, and Francesco De Mattia. "Cultural Heritage Digital Libraries on Data Grids." *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, Corfu, Greece, (2009): 469-472. http://dx.doi.org/10.1007/978-3-642-04346-8_61

[Candela et al. 2007]. Candela, Leonardo, Donatella Castelli, Pasquale Pagano, Consantino Thanos, Yannis Ioannidis, Georgia Koutrika, Seamus Ross, Hans Jörg Schek, and Heiko Schuldt. "Setting the Foundations of Digital Libraries: The DELOS Manifesto." *D-Lib Magazine*, 13, (2007). http://www.dlib.org/dlib/march07/castelli/03castelli.html

[Cantara 2006]. Cantara, Linda. "Long-Term Preservation of Digital Humanities Scholarship." *OCLC Systems & Services*, 22 (2006): 38-42.

[Campbell 2007]. Campbell, Douglas. "Identifying the Identifiers." *2007 Proceedings of the International Conference on Dublin Core and Metadata Applications*, (2007): 74-84. http://www.dcmipubs.org/ojs/index.php/pubs/article/view/34/16

[Carusi and Reimer 2010]. Carusi, Annamaria and Torsten Reimer. "Virtual Research Environment Collaborative Landscape Study: A JISC Funded Project." Joint Information Systems Committee, (January 2010). http://www.jisc.ac.uk/media/documents/publications/vrelandscapereport.pdf

[Casadio and Lambek 2005]. Casadio, Claudia and Jim Lambek. "A Computational Algebraic Approach to Latin Grammar." *Research on Language & Computation*, 3 (April 2005): 45-60. http://dx.doi.org/10.1007/s11168-005-1286-0

[Cayless 2008]. Cayless, Hugh A. "Linking Page Images to Transcriptions with SVG." *Balisage: The Markup Conference 2008*, (August 2008): 12-15. http://www.balisage.net/Proceedings/vol1/html/Cayless01/BalisageVol1-Cayless01.html

[Cayless 2009]. Cayless, Hugh A. "Image as Markup: Adding Semantics to Manuscript Images." *Digital Humanities 2009 Conference Abstracts*, (June 2009): 83-84. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf

[Cayless 2010a]. Cayless, Hugh (2010) "Digitized Manuscripts and Open Licensing." *Digital Proceedings of the Lawrence J. Schoenberg Symposium on Manuscript Studies in the Digital Age*, 2 (1), Article 7, (2010). http://repository.upenn.edu/ljsproceedings/vol2/iss1/7

[Cayless 2010b]. Cayless, Hugh A. "*Ktêma es aiei*: Digital Permanence from an Ancient Perspective." In *Digital Research in the Study of Classical Antiquity* (eds. Gabriel Bodard and Simon Mahony). Burlington, VT: Ashgate Publishing, 2010, pp. 139-150.

[Cayless 2010c]. Cayless, Hugh A. "Making a New Numbers Server for Papyri.info." *Scriptio Continua*, (March 2, 2010). **http://philomousos.blogspot.com/2010/03/making-new-numbers-server-for.html**

[Cayless et al. 2009]. Cayless, Hugh, Charlotte Roueché, Tom Elliott, and Gabriel Bodard. "Epigraphy in 2017." *Digital Humanities Quarterly*, 3 (January 2009). http://www.digitalhumanities.org/dhq/vol/3/1/000030.html

[Ciechomski et al. 2004]. Ciechomski, Pablo de Heras, Branislav Ulicny, Rachel Cetre and Daniel Thalmann. "A Case Study of a Virtual Audience in a Reconstruction of an Ancient Roman Odeon in Aphrodisias." *VAST 2004: The 5th International Symposium on Virtual Reality, Archaeology and Cultural Heritage*, (2004). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.3403&rep=rep1&type=pdf

[Choudhury 2008]. Choudhury, Sayeed G. "Case Study in Data Curation at Johns Hopkins University." *Library Trends*, 57 (2008): 211-220. http://muse.jhu.edu/journals/library_trends/v057/57.2.choudhury.html

[Choudhury and Stinson 2007]. Choudhury, Sayeed G. and Timothy L. Stinson. "The Virtual Observatory and the Roman de la Rose: Unexpected Relationships and the Collaborative Imperative." *Academic Commons*, (December 2007). http://www.academiccommons.org/commons/essay/VO-and-roman-de-la-rose-collaborative-imperative

[Choudhury et al. 2006]. Choudhury, Sayeed G., Tim DiLauro, Robert Ferguson, Michael Droettboom, and Ichiro Fujinaga. "Document Recognition for a Million Books." *D-Lib Magazine*, 12 (2006). http://www.dlib.org/dlib/march06/choudhury/03choudhury.html

[Ciula 2009]. Ciula, Arianna. "The Palaeographical Method Under the Light of a Digital Approach." *Kodikologie und Paläographie im digitalen Zeitalter-Codicology and Palaeography in the Digital Age*. Norderstedt: Books on Demand, 2009, pp. 219-235. Also available online at: http://kups.ub.uni-koeln.de/volltexte/2009/2971/

[Clarysse and Thompson 2006]. Clarysse, Willy and Dorothy J. Thompson. *Counting the People in Hellenistic Egypt*, *Volume 2: Historical Studies* (Cambridge Classical Studies). Cambridge: Cambridge University Press, 2006.

[Clocksin 2003]. Clocksin, William F. "Towards Automatic Transcription of Syriac Handwriting." *Proceedings of the 12th Image Analysis and Processing Conference*, (2003): 664-669. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1234126

[Cohen et al. 2004]. Cohen, Jonathan, Donald Duncan, Dean Snyder, Jerrold Cooper, Subodh Kumar, Daniel Hahn, Yuan Chen, Budirijanto Purnomo, and John Graettinger. "iClay: Digitizing Cuneiform." *VAST 2004: Proceedings of the Fifth International Symposium on Virtual Reality, Archaeology, and Cultural Heritage*, (2004): 135-143. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.73.87&rep=rep1&type=pdf

[Cohen et al. 2009]. Cohen, Dan, Neil Fraistat, Matthew Kirschenbaum, and Tom Scheinfeldt. *Tools for Data-Driven Scholarship: Past, Present and Future. A Report on the Workshop of 22-24 October 2008. Turf Valley Resort, Ellicott City, Maryland*. March 2009, Center for History and New Media, George Mason University and Maryland Institute for Technology and the Humanities. http://mith.umd.edu/tools/final-report.html

[Cohen 2010]. Cohen, Patricia. "Scholars Test Web Alternative to Peer Review." *New York Times*, August 23, 2010, Published online at: http://www.nytimes.com/2010/08/24/arts/24peer.html?_r=1

[Collins 2008]. Collins, Derek. "Mapping the Entrails: The Practice of Greek Hepatoscopy." *American Journal of Philology*, 129 (2008): 319-345.

[Connaway and Dickey 2010]. Connaway, Lynn S. and Timothy J. Dickey. *The Digital Information Seeker: Report of Findings from selected OCLC, RIN and JISC User Behaviour Projects*. Higher Education Funding Council, (March 2010).

http://www.jisc.ac.uk/media/documents/publications/reports/2010/digitalinformationseekerreport.pdf

[Crane 1991]. Crane, Gregory. "Generating and Parsing Classical Greek." *Literary & Linguistic Computing*, 6 (January 1991): 243-245. http://llc.oxfordjournals.org/cgi/content/abstract/6/4/243

[Crane 1998]. Crane, Gregory. "New Technologies for Reading: The Lexicon and the Digital Library." *The Classical World*, 91 (1998): 471-501. http://www.jstor.org/stable/4352154

[Crane 2004]. Crane, Gregory. "Classics and the Computer: an End of the History." in *A Companion to Digital Humanities*. Oxford: Blackwell Publishers, 2004, pp. 46-55. Available online at: http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-2-4&toc.depth=1&toc.id=ss1-2-4&brand=9781405103213_brand

[Crane 2005]. Crane, Gregory. "In a Digital World, No Book is an Island: Designing Electronic Primary Sources and Reference Works for the Humanities." in *Creation, Deployment and Use of Digital Information*, New Jersey: Lawrence Erlbaum Associates, 2005, pp. 11-26.

[Crane 2008]. Crane, Gregory. "Repositories, Cyberinfrastructure, and the Humanities." *EDUCAUSE Review*, 6 (November 2008). http://www.educause.edu/EDUCAUSE+Review/EDUCAUSEReviewMagazineVolume43/RepositoriesCyberinfrastructur/163269

[Crane and Jones 2006]. Crane, Gregory and Alison Jones. "Text, Information, Knowledge and the Evolving Record of Humanity." *D-Lib Magazine*, 12 (March 2006). http://www.dlib.org/dlib/march06/jones/03jones.html

[Crane, Babeu and Bamman 2007]. Crane, Gregory, Alison Babeu, and David Bamman. "eScience and the Humanities." *International Journal on Digital Libraries*, 7 (October 2007): 117-122. Preprint available at: http://hdl.handle.net/10427/42690

[Crane, Seales and Terras 2009]. Crane, Greg, Brent Seales, and Melissa Terras. "Cyberinfrastructure for Classical Philology." *Digital Humanities Quarterly*, 3 (January 2009). http://www.digitalhumanities.org/dhq/vol/3/1/000023.html#

[Crane et al. 2006]. Crane, Gregory, David Bamman, Lisa Cerrato, Alison Jones, David Mimno, Adrian Packel, David Sculley, and Gabriel Weaver. "Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries." *Proceedings of the 10th European Conference on Digital Libraries (ECDL 2006)*. September 2006, 353-366. Preprint available at: http://hdl.handle.net/10427/36131

[Crane et al. 2009a]. Crane, Gregory, Alison Babeu, David Bamman, Thomas Breuel, Lisa Cerrato, Daniel Deckers, Anke Lüdeling, David Mimno, Rashmi Singhal, David A. Smith, and Amir Zeldes. "Classics in the Million Book Library." *Digital Humanities Quarterly*, 3 (2009). http://www.digitalhumanities.org/dhq/vol/003/1/000034.html

[Crane et al. 2009b] Crane, Gregory, Alison Babeu, David Bamman, Lisa Cerrato, and Rashmi Singhal. "Tools for Thinking: ePhilology and Cyberinfrastructure." in *Working Together or Apart: Promoting the Next Generation of Digital Scholarship: Report of a Workshop Cosponsored by the Council on Library and Information Resources and The National Endowment for the Humanities*. Council on Library and Information Resources, Publication Number 145, (March 2009): 16-26. http://www.clir.org/activities/digitalscholar2/crane11_11.pdf

[Csernel and Patte 2009]. Csernel, Marc and François Patte. "Critical Edition of Sanskrit Texts." *Sanskrit Computational Linguistics*, (*Lecture Notes in Computer Science*, Volume 5402), Springer, (2009): 358-379. http://dx.doi.org/10.1007/978-3-642-00155-0_19

[Dalbello et al. 2006]. Dalbello, Marija, Irene Lopatovska, Patricia Mahony, and Nomi Ron. "Electronic Texts and the Citation System of Scholarly Journals in the Humanities: Case Studies of Citation Practices in the Fields of Classical Studies and English Literature." *Proceedings of Libraries in the Digital Age (LIDA)*, (2006). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.7025&rep=rep1&type=pdf

[Dallas and Doorn 2009]. Dallas, Costis and Peter Doorn. "Report on the Workshop on Digital Curation in the Human Sciences at ECDL 2009: Corfu, 30 September - 1 October 2009." *D-Lib Magazine*, 15 (November 2009). http://www.dlib.org/dlib/november09/dallas/11dallas.html

[D'Andrea and Niccolucci 2008]. D'Andrea, Andrea and Franco Niccolucci. "Mapping, Embedding and Extending: Pathways to Semantic Interoperability The Case of Numismatic Collections." *Fifth European Semantic Web Conference Workshop: SIEDL 2008-Semantic Interoperability in the European Digital Library*, (2008): 63-76. http://image.ntua.gr/swamm2006/SIEDLproceedings.pdf#page=69

[Deckers et al. 2009]. Deckers, Daniel, Lutz Koll, and Cristina Vertan. "Representation and Encoding of Heterogeneous Data in a Web Based Research Environment for Manuscript and Textual Studies." *Kodikologie und Paläographie im digitalen Zeitalter-Codicology and Palaeography in the Digital Age*. Norderstedt: Books on Demand, 2009. Also available online at: http://kups.ub.uni-koeln.de/volltexte/2009/2962/

[de Jong 2009]. de Jong, Franciscka. "Invited Talk: NLP and the Humanities: The Revival of an Old Liaison." *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, Athens, Greece, (2009): 10-15. http://www.aclweb.org/anthology-new/E/E09/E09-1002.pdf

[de la Flor et al. 2010]. Flor, Grace de la, Paul Luff, Marina Jirotka, John Pybus, Ruth Kirkham, and Annamaria Carusi. "The Case of the Disappearing Ox: Seeing Through Digital Images to an Analysis of Ancient Texts." *CHI '10: Proceedings of the 28th international Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, (2010): 473-482. http://dx.doi.org/10.1145/1753326.1753397

[Dekhtyar et al. 2005]. Dekhtyar, Alex, Ionut E. Iacob, Jerzy W. Jaromczyk, Kevin Kiernan, Neil Moore, and Dorothy Carr Porter. "Support for XML Markup of Image-Based Electronic Editions." *International Journal on Digital Libraries*, 6 (2006): 55-69. Preprint available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.2471&rep=rep1&type=pdf

[Deufert et al. 2010]. Deufert, Marcus, Judith Blumenstein, Andreas Trebesius, Stefan Beyer, and Marco Büchler. "Objective Detection of Plautus' Rules by Computer Support." *Digital Humanities 2010 Conference Abstracts*, (July 2010): 126-128. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/book-final.pdf

[Dik and Whaling 2009]. Dik, Helma and Richard Whaling. "Implementing Greek Morphology." *Digital Humanities 2009 Conference Abstracts*, (June 2009): 338-339. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf

[Dimitriadis et al. 2006]. Dimitriadis, Alexis, Marc Kemps-Snijders, Peter Wittenburg, M. Everaert, and S. Levinson. "Towards a Linguist's Workbench Supporting eScience Methods." *E-SCIENCE '06: Proceedings of the Second IEEE International Conference on e-Science and Grid Computing*. IEEE Computer Society, (2006): 131-. http://www.lat-mpi.eu/papers/papers-2006/escience-sketch-final2.pdf/view

[Dogan and Scharsky 2008]. Dogan, Zeki and Alfred Scharsky. "Virtual Unification of the Earliest Christian Bible: Digitisation, Transcription, Translation and Physical Description of the Codex Sinaiticus." *ECDL 2008: Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries*, (2008): 221-226. http://dx.doi.org/10.1007/978-3-540-87599-4_22

[Doumat et al. 2008]. Doumat, Reim, Elöd E. Zsigmond, Jean M. Pinon, and Emese Csiszar. "Online Ancient Documents: Armarius." *DocEng '08: Proceeding of the Eighth ACM Symposium on Document Engineering*. New York, NY, USA: ACM, (2008): 127-130. http://dx.doi.org/10.1145/1410140.1410167

[Doerr and Iorizzo 2008]. Doerr, Martin and Dolores Iorizzo. "The Dream of a Global Knowledge Network—A New Approach." *Journal on Computing and Cultural Heritage*, 1 (June 2008): 1-23. http://dx.doi.org/10.1145/1367080.1367085

[Dué and Ebbott 2009]. Dué, Casey and Mary Ebbott. "Digital Criticism: Editorial Standards for the Homer Multitext." *Digital Humanities Quarterly*, 3 (January 2009). http://www.digitalhumanities.org/dhq/vol/3/1/000029.html#

[Dunn 2009]. Dunn, Stuart. "Dealing with the Complexity Deluge: VREs in the Arts and Humanities." *Library Hi Tech*, 27 (2009): 205-216. http://dx.doi.org/10.1108/07378830910968164

[Dunn 2010]. Dunn, Stuart. "Space as an Artefact: A Perspective on 'Neogeography' from the Digital Humanities." in *Digital Research in the Study of Classical Antiquity* (eds Gabriel Bodard and Simon Mahony). Ashgate Publishing, 2010, pp. 53-69

[Dutschke 2008]. Dutschke, Consuelo W. "Digital Scriptorium: Ten Years Young, and Working on Survival." *Storicamente*, 4, (2008). http://www.storicamente.org/02_tecnostoria/filologia_digitale/dutschke.html#_ftn1

[Ebeling 2007]. Ebeling, Jarle. "The Electronic Text Corpus of Sumerian Literature." *Corpora*, 2 (2007): 111-120. http://www.eupjournals.com/doi/abs/10.3366/cor.2007.2.1.111

[Eder 2007]. Eder, Maciej. "How Rhythmical is Hexameter: A Statistical Approach to Ancient Epic Poetry." *Digital Humanities 2007 Conference Abstracts*, (May 2007). http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=137

[Edmond and Schreibman 2010]. Edmond, Jennifer and Susan Schreibman. "European Elephants in the Room (Are They the Ones With the Bigger or Smaller Ears?)" *Online Humanities Scholarship: The Shape of Things to Come*, Rice University Press, (March 2010). http://rup.rice.edu/cnx_content/shape/m34307.html

[Edwards et al. 2004]. Edwards, Jaety, Yee W. Teh, David Forsyth, Roger Bock, and Michael Maire. "Making Latin Manuscripts Searchable Using gHMMs." *NIPS*, (2004). http://books.nips.cc/papers/files/nips17/NIPS2004_0550.pdf

[Eglin et al. 2006]. Eglin, Véronique, Frank Lebourgeois, S. Bres, Hubert Emptoz, Yann Leydier, Ikram Moalla, and F. Drira. "Computer Assistance for Digital Libraries: Contributions to Middle-Ages and Authors' Manuscripts Exploitation and Enrichment." *DIAL 2006: Second International Conference on Document Image Analysis for Libraries*, (2006). http://dx.doi.org/10.1109/DIAL.2006.9

[Elliott 2008]. Elliott, Michelle. "Introducing the "Digital Archaeological Record."" , (2008). http://www.tdar.org/confluence/download/attachments/131075/tDAR_Instructions.pdf?version=1&modificationDate=1218787124032

[Elliott and Gillies 2009a]. Elliott, Tom and Sean Gillies. "Data and Code for Ancient Geography: Shared Effort Across Projects and Disciplines." *Digital Humanities 2009 Conference Abstracts*, (June 2009): 4-6. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf

[Elliott and Gillies 2009b]. Elliott, Tom and Sean Gillies. "Digital Geography and Classics." *Digital Humanities Quarterly*, 3 (January 2009). http://www.digitalhumanities.org/dhq/vol/3/1/000031.html

[Emery and Toth 2009] Emery, Doug and Michael B. Toth. "Integrating Images and Text with Common Data and Metadata Standards in the Archimedes Palimpsest." *Digital Humanities 2009 Conference Abstracts*, (June 2009): 281-283.
http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf

[Ernst-Gerlach and Crane 2008]. Ernst-Gerlach, Andrea and Gregory Crane. "Identifying Quotations in Reference Works and Primary Materials." *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries*, (2008): 78-87. http://dx.doi.org/10.1007/978-3-540-87599-4_9

[Feijen et al. 2007]. Feijen, Martin, Wolfram Horstmann, Paolo Manghi, Mary Robinson, and Rosemary Russell. "DRIVER: Building the Network for Accessing Digital Repositories across Europe." *Ariadne*, 53 (October 2007). http://www.ariadne.ac.uk/issue53/feijen-et-al/

[Feraudi-Gruénais 2010]. Feraudi-Gruénais, Francisca, ed. *Latin on Stone: Epigraphic Research and Electronic Archives* (Roman Studies: Interdisciplinary Approaches). Lahnam: Lexington Books, 2010. http://www.worldcat.org/oclc/526091486

[Finkel and Stump 2009]. Finkel, Raphael and Gregory Stump. "What Your Teacher Told You is True: Latin Verbs Have Four Principal Parts." *Digital Humanities Quarterly*, 3 (January 2009). http://www.digitalhumanities.org/dhq/vol/3/1/000032.html

[Flanders 2009]. Flanders, David F. *Fedorazon: Final Report*. Joint Information Systems Committee, (November 2009). http://ie-repository.jisc.ac.uk/426/

[Flaten 2009]. Flaten, Arne R. "The Ashes2Art Project: Digital Models of Fourth-Century BCE Delphi, Greece." *Visual Resources: An International Journal of Documentation*, 25 (December 2009): 345-362. http://dx.doi.org/10.1080/01973760903331783

[Forstall and Scheirer 2009]. Forstall, Christopher W. and W. J. Scheirer. "Features from Frequency: Authorship and Stylistic Analysis Using Repetitive Sound." *DHCS 2009-Chicago Colloquium on Digital Humanities and Computer Science*. November 2009. http://lingcog.iit.edu/%7Eargamon/DHCS09-Abstracts/Forstall.pdf

[Fraser 2005]. Fraser, Michael. "Virtual Research Environments: Overview and Activity." *Ariadne*, 44 (July 2005). http://www.ariadne.ac.uk/issue44/fraser/

[Fraser 2008]. Fraser, Bruce L. "Beyond Definition: Organising Semantic Information in Bilingual Dictionaries." *International Journal of Lexicography*, 21 (March 2008): 69-93. http://dx.doi.org/10.1093/ijl/ecn002

[Friedlander 2009]. Friedlander, Amy. "Asking Questions and Building a Research Agenda for Digital Scholarship." in *Working Together or Apart: Promoting the Next Generation of Digital Scholarship: Report of a Workshop Cosponsored by the Council on Library and Information Resources and The National Endowment for the Humanities*. Council on Library and Information Resources, Publication Number 145 (March 2009): 1-15.

http://www.clir.org/pubs/reports/pub145/pub145.pdf

[Fulford et al. 2010]. Fulford, Michael G., Emma J. O'Riordan, Amanda Clarke, and Michael Rains. "Silchester Roman Town: Developing Virtual Research Practice 1997-2008." in *Digital Research in the Study of Classical Antiquity* (eds. Gabriel Bodard and Simon Mahony). Ashgate Publishing, 2010, pp. 16-34.

[Fusi 2008]. Fusi, Daniele. "An Expert System for the Classical Languages: Metrical Analysis Components." (2008). http://fusisoft.it/Doc/ActaVenezia.pdf

[Gelernter and Lesk 2008]. Gelernter, Judith and Michael E. Lesk. "Traditional Resources Help Interpret Texts." *BooksOnline '08: Proceeding of the 2008 ACM Workshop on Research Advances in Large Digital Book Repositories*. New York, NY, USA: ACM, (2008): 17-20. http://dx.doi.org/10.1145/1458412.1458418

[Gietz et al. 2006]. Gietz, Peter, Andreas Aschenbrenner, Stefan Budenbender, Fotis Jannidis, Marc W. Küster, Christoph Ludwig, Wolfgang Pempe, Thorsten Vitt, Werner Wegstein, and Andrea Zielinski. "TextGrid and eHumanities." *E-SCIENCE '06: Proceedings of the Second IEEE International Conference on e-Science and Grid Computing*. Washington, DC, USA: IEEE Computer Society, (2006). http://dx.doi.org/10.1109/E-SCIENCE.2006.133

[Gill 2009]. Gill, Alyson A. "Digitizing the Past: Charting New Courses in the Modeling of Virtual Landscapes." *Visual Resources: An International Journal of Documentation*, 25 (2009): 313-332.

[Goudriaan et al. 1995]. Goudriaan, Koen, Kees Mandemakers, Jogchum Reitsma and Peter Stabel, eds. *Prosopography and Computer: Contributions of Mediaevalists and Modernists on the Use of Computer in Historical Research*. Leuven, 1995.

[Graham and Ruffini 2007]. Graham, Shawn and Giovanni Ruffini. "Network Analysis and Greco-Roman Prosopography." in *Prosopography Approaches and Applications: A Handbook*. Ed. K. S. B. Keats-Rohan Oxford: Unit for Prosopographical Research, Linacre College, University of Oxford, 2007, pp. 325-336.

[Green and Roy 2008] Green, David and Michael Roy. "Things to Do While Waiting for the Future to Happen: Building Cyberinfrastructure for the Liberal Arts." *EDUCAUSE Review*, 43 (July 2008). http://connect.educause.edu/display/46969

[Gruber 2009]. Gruber, Ethan. "Encoded Archival Description for Numismatic Collections." Presentation made at *Computer Applications and Quantitative Methods in Archaeology*, (March 2009). http://coins.lib.virginia.edu/documentation/caa2009.pdf

[Guidi et al. 2006]. Guidi, Gabriele, Bernard Frischer, Michele Russo, Alessandro Spinetti, Luca Carosso, and Laura Micoli. "Three-Dimensional Acquisition of Large and Detailed Cultural Heritage Objects." *Machine Vision and Applications*, 17 (December 2006): 349-360. http://dx.doi.org/10.1007/s00138-006-0029-z

[Hahn et al. 2006]. Hahn, Daniel V., Donald D. Duncan, Kevin C. Baldwin, Jonathon D. Cohen, and Budirijanto Purnomo. "Digital Hammurabi: Design and Development of a 3D Scanner for Cuneiform Tablets." *Proceedings of SPIE, Vol 6056: Three-Dimensional Image Capture and Applications VII*, (January 2006). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.73.1965

[Hanson 2001]. Hanson, Ann E. "Papyrology: Minding Other People's Business." *Transactions of the American Philological Association*, 131 (2001): 297-313. http://www.jstor.org/stable/20140974

[Hardwick 2000]. Hardwick, Lorna. "Electrifying the Canon: The Impact of Computing on Classical Studies." *Computers and the Humanities*, 34 (August 2000): 279-295. http://dx.doi.org/10.1023/A:1002089109613

[Harley et al. 2006a]. Harley, Diane, Jonathan Henke, Shannon Lawrence, Ian Miller, Irene Perciali, David Nasatir, Charis Kaskiris, Cara Bautista. *Use and Users of Digital Resources: A Focus on Undergraduate Education in the Humanities and Social Sciences*. Center for Studies in Higher Education, (April 5, 2006). http://www.escholarship.org/uc/item/8c43w24h

[Harley et al. 2006b]. Harley, Diane, Jonathan Henke, and Shannon Lawrence. "Why Study Users? An Environmental Scan of Use and Users of Digital Resources in Humanities and Social Sciences Undergraduate Education." Research & Occasional Paper Series: CSHE.15.06. Center for Studies in Higher Education, (September 2006). http://www.escholarship.org/uc/item/61g3s91k

[Harley et al. 2010]. Harley, Diane, Sophia K. Acord, Sarah Earl-Novell, Shannon Lawrence, and C. Judson King. *eScholarship: Assessing the Future Landscape of Scholarly Communication*. University of California-Berkeley: Center for Studies in Higher Education, (January 2010). http://escholarship.org/uc/cshe_fs

[Heath 2010]. Heath, Sebastian. "Diversity and Reuse of Digital Resources for Ancient Mediterranean Material Culture." in *Digital Research in the Study of Classical Antiquity* (eds. Gabriel Bodard and Simon Mahony). Ashgate Publishing Company, 2010, pp. 35-52. http://sebastianheath.com/files/HeathS2010-DigitalResearch.pdf

[Hedges 2009]. Hedges, Mark. "Grid-Enabling Humanities Datasets." *Digital Humanities Quarterly*, 3 (4), (2009). http://www.digitalhumanities.org/dhq/vol/3/4/000078/000078.html#

[Hellwig 2007]. Hellwig, Oliver. "SanskritTagger: A Stochastic Lexical and POS Tagger for Sanskrit." *Sanskrit Computational Linguistics*, (*Lectures Notes in Computer Science*, Volume 5402), Spring (2009): 266-277. http://dx.doi.org/10.1007/978-3-642-00155-0_11

[Hellwig 2010]. Hellwig, Oliver. "Etymological Trends in the Sanskrit Vocabulary." *Literary & Linguistic Computing*, 25 (1) (October 2009): 105-118. http://dx.doi.org/10.1093/llc/fqp034

[Hillen 2007]. Hillen, Michael, (translated by Kathleen M. Coleman). "Finishing the TLL in the Digital Age: Opportunities, Challenges, Risks." *Transactions of the American Philological Association*, 137 (2007): 491-495.

[Hilse and Kothe 2006]. Hilse, Hans Werner and Jochen Kothe. *Implementing Persistent Identifiers.* Research and Development Department of the Goettingen State and University Library, (November 2006). http://www.knaw.nl/ecpa/publ/pdf/2732.pdf

[Honigman 2004]. Honigman, Sylvie. "Abraham in Egypt: Hebrew and Jewish-Aramaic Names in Egypt and Judaea in Hellenistic and Early Roman Times." *Zeitschrift für Papyrologie und Epigraphik*, 146 (2004): 279-297.

[Huet 2004]. Huet, Gérard. "Design of a Lexical Database for Sanskrit." *ElectricDict '04: Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*. Morristown, NJ, USA: Association for Computational Linguistics, (2004), 8-14. http://portal.acm.org/citation.cfm?id=1610042.1610045

[Hunt et al. 2005]. Hunt, Leta, Marilyn Lundberg, and Bruce Zuckerman. "InscriptiFact: A Virtual Archive of Ancient Inscriptions from the Near East." *International Journal on Digital Libraries*, 5 (May 2005): 153-166. http://www.inscriptifact.com/news/JDL.pdf

[IFLA 1998] International Federation of Library Associations (IFLA). *Functional Requirements for Bibliographic Records: Final Report, Volume 19 of UBCIM Publications-New Series*. München: K.G.Saur. http://www.ifla.org/VII/s13/frbr/frbr.pdf.

[Isaksen 2008]. Isaksen, Leif. "The Application Of Network Analysis To Ancient Transport Geography: A Case Study Of Roman Baetica." *Digital Medievalist*, (2008). http://www.digitalmedievalist.org/journal/4/isaksen/

[Isaksen 2009]. Isaksen, Leif. "Augmenting Epigraphy." Paper presented at "Object Artefact Script Workshop." (October 8-9 2009). http://wiki.esi.ac.uk/Object_Artefact_Script_Isaksen

[Isaksen 2010]. Isaksen, Leif. "Reading Between the Lines: Unearthing Structure in Ptolemy's Geography." *Digital Classicist/ICS Work in Progress Seminar*, (June 2010). http://www.digitalclassicist.org/wip/wip2010-01li.pdf

[Jackson et al. 2009]. Jackson, Mike, Mario Antonioletti, Alastair Hume, Tobias Blanke, Gabriel Bodard, Mark Hedges, and Shrija Rajbhandari. "Building Bridges Between Islands of Data - An Investigation into Distributed Data Management in the Humanities." *e-Science '09: Fifth IEEE International Conference on e-Science*, (December 2009): 33-39. http://dx.doi.org/10.1109/e-Science.2009.13

[Jaworski 2008]. Jaworski, Wojciech. "Contents Modelling of Neo-Sumerian Ur III Economic Text Corpus." *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, (2008): 369-376. http://www.aclweb.org/anthology/C08-1047

[Jeffrey et al. 2009a]. Jeffrey, Stuart, Julian Richards, Fabio Ciravegna, Stewart Waller, Sam Chapman, and Ziqui Zhang. "The Archaeotools Project: Faceted Classification And Natural Language Processing in an Archaeological Context." *Philosophical Transactions of the Royal Society*, A 367 (June 2009): 2507-2519. http://rsta.royalsocietypublishing.org/content/367/1897/2507?rss=1.abstract

[Jeffrey et al. 2009b]. Jeffrey, Stuart, Julian Richards, Fabio Ciravegna, Stewart Waller, Sam Chapman, and Ziqui Zhang. "Integrating Archaeological Literature Into Resource Discovery Interfaces Using Natural Language Processing And Name Authority Services." *5th IEEE International Conference on E-Science Workshops*, IEEE, (December 2009): 184-187. http://dx.doi.org/10.1109/ESCIW.2009.5407967

[Jones 2010]. Jones, Charles E. "Going AWOL (ancientworldonline.blogspot.com/): Thoughts On Developing a Tool for the Organization and Discovery of Open Access Scholarly Resources for the Study of the Ancient World." *CSA Newsletter*, XXII (January 2010). http://www.etana.org/abzu/abzu-displayentry.pl?RC=21784

[Kainz 2009]. Kainz, Chad J. "Bamboo: Another Part of the Jigsaw." Presentation at Digital Humanities, (June 2009). http://projectbamboo.org/files/presentations/0906_Bamboo_Jigsaw.pdf

[Kampel and Zaharieva 2008]. Kampel, Martin and Maia Zaharieva. "Recognizing Ancient Coins Based on Local Features." *Advances in Visual Computing* (*Lecture Notes in Computer Science*, Vol. 5358), (2008): 11-22. http://dx.doi.org/10.1007/978-3-540-89639-5_2

[Kansa et al. 2007]. Kansa, Sarah W., Eric C. Kansa, and Jason M. Schultz. "An Open Context for Near Eastern Archaeology." *Near Eastern Archaeology*, 70 (December 2007): 188-194. http://www.alexandriaarchive.org/publications/KansaKansaSchultz_NEADec07.pdf

[Knight and Pennock 2008]. Knight, Gareth and Maureen Pennock. "Data Without Meaning: Establishing the Significant Properties of Digital Research." *iPRES 2008: The Fifth International Conference on Preservation of Digital Objects*, (September 2008). http://www.bl.uk/ipres2008/presentations_day1/16_Knight.pdf

[Knoll et al. 2009]. Knoll, Adolf, Tomáš Psohlavec, Stanislav Psohlavec, and Zdeněk Uhlíř. "Creation of an International Digital Library of Manuscripts: Seamless Access to Data from Heterogeneous Resources (ENRICH Project)." *ELPUB 2009: 13th International Conference on Electronic Publishing: Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies*, (June 2009): 335-347. http://conferences.aepic.it/index.php/elpub/elpub2009/paper/view/71/30

[Koller et al. 2009]. Koller, David, Bernard Frischer, and Greg Humphreys. "Research Challenges for Digital Archives of 3D Cultural Heritage Models." *Journal on Computing and Cultural Heritage*, 2 (3), (2009): 1-17. http://dx.doi.org/10.1145/1658346.1658347

[Kretzschmar 2009]. Kretzschmar, William A. "Large-Scale Humanities Computing Projects: Snakes Eating Tails, or Every End is a New Beginning?" *Digital Humanities Quarterly*, 3 (June 2009). http://www.digitalhumanities.org/dhq/vol/3/2/000038.html

[Krmnicek and Probst 2010]. Krmnicek, Stefan and Peter Probst. "Open Access, Classical Studies and Publication by Postgraduate Researchers" *Archaeolog*, (May 22, 2010). http://traumwerk.stanford.edu/archaeolog/2010/05/open_access_classical_studies.html

[Kumar et al. 2003]. Kumar, Subodh, Dean Snyder, Donald Duncan, Jonathan Cohen, and Jerry Cooper. "Digital Preservation of Ancient Cuneiform Tablets Using 3D-Scanning." *Proceedings of the Fourth International Conference on 3-D Digital Imaging and Modeling* (3DIM'03), (October 2003): 326-333. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.58.9706&rep=rep1&type=pdf

[Kurtz et al. 2009]. Kurtz, Donna, Greg Parker, David Shotton, Graham Klyne, Florian Schroff, Andrew Zisserman, and Yorick Wilks. "CLAROS - Bringing Classical Art to a Global Public." *Fifth IEEE International Conference on e-Science'09*, (December 2009): 20-27. http://www.clarosnet.org/PID1023719.pdf

[Lee 2007]. Lee, John. "A Computational Model of Text Reuse in Ancient Literary Texts." *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic: Association for Computational Linguistics, (2007): 472-479. http://acl.ldc.upenn.edu/P/P07/P07-1060.pdf

[Lee 2008]. Lee, John. "A Nearest-Neighbor Approach to the Automatic Analysis of Ancient Greek Morphology." *CoNLL '08: Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Morristown, NJ, USA: Association for Computational Linguistics, (2008): 127-134. http://www.aclweb.org/anthology/W/W08/W08-2117.pdf

[Leydier et al. 2007]. Leydier, Yann, Frank LeBourgeois, and Hubert Emptoz. "Text Search for Medieval Manuscript Images." *Pattern Recognition*, 40 (December 2007): 3552-3567. http://dx.doi.org/10.1016/j.patcog.2007.04.024

[Leydier et al. 2009]. Leydier, Yann, Asma Ouji, Frank LeBourgeois, and Hubert Emptoz. "Towards An Omnilingual Word Retrieval System For Ancient Manuscripts." *Pattern Recognition*, 42, (September 2009): 2089-2105. http://dx.doi.org/10.1016/j.patcog.2009.01.026

[Lock 2003]. Lock, Gary. *Using Computers In Archaeology: Towards Virtual Pasts*. London, UK: Routledge, 2003.

[Lockyear 2007]. Lockyear, Kris. "Where Do We Go From Here? Recording and Analysing Roman Coins from Archaeological Excavations." *Britannia*, (November 2007): 211-224.

[Lüdeling and Zeldes 2007]. Lüdeling, Anke and Amir Zeldes. “Three Views on Corpora: Corpus Linguistics, Literary Computing, and Computational Linguistics.” *Jahrbuch für Computerphilologie*, 9 (2007): 149-178. http://computerphilologie.tu-darmstadt.de/jg07/luedzeldes.html

[Ludwig and Küster 2008]. Ludwig, Christoph and Marc Wilhelm Küster. “Digital Ecosystems of eHumanities Resources and Services.” *DEST 2008: 2nd IEEE International Conference on Digital Ecosystems and Technologies*, (2008): 476-481. http://dx.doi.org/10.1109/DEST.2008.4635178

[Lynch 2002]. Lynch, Clifford. “Digital Collections, Digital Libraries and the Digitization of Cultural Heritage Information.” *First Monday*, 7 (5), (May 2002). http://www.firstmonday.org/issues/issue7_5/lynch/

[MacMahon 2006]. MacMahon, Cary. “Using and Sharing Online Resources in History, Classics and Archaeology.” The Higher Education Academy--Subject Centre for History, Classics and Archaeology, (2006). http://www.heacademy.ac.uk/assets/hca/documents/UsingandSharingOnlineResourcesHCA.pdf

[Mahoney 2009]. Mahoney, Anne. “Tachypaedia Byzantina: The Suda On Line as Collaborative Encyclopedia.” *Digital Humanities Quarterly*, 3 (January 2009). http://www.digitalhumanities.org/dhq/vol/3/1/000025.html#

[Mahony 2006]. Mahony, Simon. “New Tools for Collaborative Research: the Example of the Digital Classicist Wiki.” *CLiP 2006: Literatures, Languages and Cultural Heritage in a Digital World*, (2006). http://www.cch.kcl.ac.uk/clip2006/content/abstracts/paper31.html

[Mahony and Bodard 2010]. Mahony, Simon and Gabriel Bodard. “Introduction.” in *Digital Research in the Study of Classical Antiquity* (eds. G. Bodard and S. Mahony). Ashgate Publishing, 2010, pg. 1-14.

[Mallon 2006]. Mallon, Adrian. “eLingua Latina: Designing a Classical-Language E-Learning Resource.” *Computer Assisted Language Learning*, 19 (2006): 373-387.

[Marchionini and Crane 1994]. Marchionini, Gary and Gregory Crane. “Evaluating Hypermedia and Learning: Methods and Results from the Perseus Project.” *ACM Transactions on Information Systems (TOIS)*, 12 (January 1994): 5-34. http://dx.doi.org/10.1145/174608.174609

[Marek 2009]. Marek, Jindřich. “Creating Document Management System for the Digital Library of Manuscripts: M-Tool and M-Can for Manuscriptorium.” Presentation given at *TEI 2009 Members' Meeting*, (November 2009). http://www.lib.umich.edu/spo/teimeeting09/files/TEI_MM_2009_MSS_SIG_Marek.pdf

[Marinai 2009]. Marinai, Simone. “Text Retrieval from Early Printed Books.” *AND '09: Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*. New York, NY, USA: ACM, (2009): 33-40. http://dx.doi.org/10.1145/1568296.1568304

[Marshall 2008]. Marshall, Catherine C. “From Writing and Analysis to the Repository: Taking the Scholars' Perspective on Scholarly Archiving.” *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*. New York, NY, USA: ACM, (2008): 251-260. http://dx.doi.org/10.1145/1378889.1378930

[Martinez-Uribe and Macdonald 2009]. Martinez-Uribe, Luis and Stuart Macdonald. “User Engagement in Research Data Curation.” *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, Corfu, Greece, Springer, (2009): 309-314. http://dx.doi.org/10.1007/978-3-642-04346-8_30

[Mathisen 1988]. Mathisen, Ralph W. “Medieval Prosopography and Computers: Theoretical and

Methodological Considerations" *Medieval Prosopography*, 9 (1988): 73–128.

[Mathisen 2007]. Mathisen, Ralph W. "Where are all the PDBs?: The Creation of Prosopographical Databases for the Ancient and Medieval Worlds." *Prosopography Approaches and Applications: A Handbook*. Ed. K. S. B. Keats-Rohan. Oxford: Unit for Prosopographical Research, Linacre College, University of Oxford, 2007, pp. 95-126. Also available online at http://prosopography.modhist.ox.ac.uk/images/04%20Mathisen%20pdf.pdf

[Matthews and Rahtz 2008]. Matthews, Elaine and Sebastian Rahtz. "The Lexicon of Greek Personal Names and Classical Web Services." *Digital Classicist/ICS Work in Progress Seminar*, (June 2008). http://www.digitalclassicist.org/wip/wip2008-01emsr.pdf

[Meckseper and Warwick 2003]. Meckseper, Christiane and Claire Warwick. "The Publication of Archaeological Excavation Reports Using XML." *Literary & Linguistic Computing*, 18 (April 2003): 63-75. http://dx.doi.org/10.1093/llc/18.1.63

[McDonough 2009]. McDonough, Jerome P. "Aligning METS with the OAI-ORE Data Model." *JCDL '09: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries.* New York, NY, USA: ACM, 2009, 323-330. Preprint available at http://hdl.handle.net/2142/10744

[McDonough 1959]. McDonough, James. "Computers and Classics." *Classical World*, 53 (2), (November 1959): 44-50. http://www.jstor.org/stable/4344244

[McGillivray and Passarotti 2009]. McGillivray, Barbara and Marco Passarotti. "The Development of the Index Thomisticus Treebank Valency Lexicon." *LaTeCH-SHELT&R '09: Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education.* Morristown, NJ, USA: Association for Computational Linguistics, (2009): 43-50. http://portal.acm.org/citation.cfm?id=1642049.1642055

[McGovern 2007]. McGovern, Nancy Y. "A Digital Decade: Where Have We Been and Where Are We Going in Digital Preservation?" *RLG DigiNews*, 11 (April 2007). http://hdl.handle.net/2027.42/60441

[McManus and Rubino 2003]. McManus, Barbara F. and Carl A. Rubino. "Classics and Internet Technology." *The American Journal of Philology*, 124 (2003): 601-608. http://www.jstor.org/stable/1561793?cookieSet=1

[Meyer et al. 2009]. Meyer, Eric T., Kathryn Eccles, and Christine Madsen. "Digitisation as e-Research Infrastructure: Access to Materials and Research Capabilities in the Humanities." Oxford Internet Institute, Oxford University, (2009). http://www.ncess.ac.uk/resources/content/papers/Meyer(2).pdf

[Mimno 2009]. Mimno, David. "Reconstructing Pompeian Households." *Applications of Topic Models Workshop-NIPS 2009*, (2009). http://www.cs.umass.edu/~mimno/papers/pompeii.pdf

[Moalla et al. 2006]. Moalla, Ikram, Frank Lebourgeois, Hubert Emptoz, and Adel Alimi. "Contribution to the Discrimination of the Medieval Manuscript Texts: Application in the Palaeography." *Document Analysis Systems VII*, (2006): 25-37. http://dx.doi.org/10.1007/11669487_3

[Monella 2008]. Monella, Paolo. "Towards a Digital Model to Edit the Different Paratextuality Levels within a Textual Tradition." *Digital Medievalist*, (March 2008). http://www.digitalmedievalist.org/journal/4/monella/

[Montanari 2004]. Montanari, Franco. "Electronic Tools for Classical Philology: The Aristarchus Project On Line." *Zbornik Matice srpske za klasične studije*, (2004): 155-160. http://scindeks-clanci.nb.rs/data/pdf/1450-6998/2004/1450-69980406155M.pdf

[Mueller and Lee 2004]. Mueller, Katja and William Lee. "From Mess to Matrix and Beyond: Estimating the Size of Settlements in the Ptolemaic Fayum/Egypt." *Journal of Archaeological Science*, 32 (January 2005): 59-67. http://dx.doi.org/10.1016/j.jas.2004.06.007

[Nagy 2010]. Nagy, Gregory. "Homer Multitext project." *Online Humanities Scholarship: The Shape of Things to Come*, Rice University Press, (March 2010). http://rup.rice.edu/cnx_content/shape/m34314.html

[Nguyen and Shilton 2008]. Nguyen, Lilly and Katie Shilton. "Tools for Humanists Project Final Report (Appendix F)." in Zorich, Diane. *A Survey of Digital Humanities Centers in the United States*. Council on Library and Information Resources, Publication Number 143, (2008). http://www.clir.org/pubs/reports/pub143/appendf.html

[Nichols 2009]. Nichols, Stephen. "Time to Change Our Thinking: Dismantling the Silo Model of Digital Scholarship." *Ariadne*, 58 (January 2009). http://www.ariadne.ac.uk/issue58/nichols/

[NSF 2008]. National Science Foundation. *Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation*. Interim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, (December 2008). http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf

[Ntzios et al. 2007]. Ntzios, Kosta, Basilios Gatos, Ioannis Pratikakis, T. Konidaris, and Stravros Perantonis. "An Old Greek Handwritten OCR System Based on an Efficient Segmentation-Free Approach." *International Journal on Document Analysis and Recognition*, 9 (April 2007): 179-192.

[Ober et al. 2007]. Ober, Josiah, Walter Scheidel, Brent D. Shaw, and Donna Sanclemente. "Toward Open Access in Ancient Studies: The Princeton-Stanford Working Papers in Classics." *Hesperia*, 76 (2007): 229-242. Preprint available at http://www.princeton.edu/~pswpc/pdfs/ober/020702.pdf

[OKell et al. 2010]. OKell, Eleanor, Dejan Ljubojevic, and Cary MacMahon. "Creating a Generative Learning Object (GLO): Working in an 'Ill-Structured' Environment and Getting Students to Think." in *Digital Research in the Study of Classical Antiquity* (eds. Gabriel Bodard and Simon Mahony). Burlington, VT: Ashgate Publishing, 2010, pp. 151-170.

[Olsen et al. 2009]. Olsen, Henriette Roued, Segolene M. Tarte, Melissa Terras, J. M. Brady, and Alan K. Bowman. "Towards an Interpretation Support System for Reading Ancient Documents." *Digital Humanities Abstracts 2009*, (June 2009): 237-239. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf

[Palmer et al. 2009]. Palmer, Carole L., Lauren C. Teffeau, and Carrie M. Pirmann. *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development*. OCLC Research, (January 2009). http://www.oclc.org/programs/publications/reports/2009-02.pdf

[Panagopoulos et al. 2008]. Panagopoulos, Michail, Constantin Papaodysseus, Panayiotis Rousopoulos, Dimitra Dafi, and Stephen Tracy. "Automatic Writer Identification of Ancient Greek Inscriptions." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31 (August 2008): 1404-1414. http://dx.doi.org/10.1109/TPAMI.2008.201

[Pappelau and Belton 2009]. Pappelau, Christine and Graham Belton. "Roman Spolia in 3D: High Resolution Leica Laserscanner Meets Ancient Building Structures." *Digital Classicist-Works in Progress Seminar*, (July 2009). http://www.digitalclassicist.org/wip/wip2009-07cp.pdf

[Pettersen et al. 2008]. Pettersen, Oystein, Nicole Bordes, Sean Ulm, David Gwynne, Terry Simmich, and Bernard Pailthorpe. "Grid Services for E-Archaeology." *AusGrid '08: Proceedings of the sixth Australasian workshop on Grid Computing and E-Research.* Darlinghurst, Australia, Australia: Australian Computer Society, Inc., (2008): 17-25. http://portal.acm.org/citation.cfm?id=1386844

[Ploeger et al. 2009]. Ploeger, Lieke, Yola Park, Jeanna N.R. Gaviria, Clemens Neudecker, Fedor Bochow, and Michael Day. "'IMPACT Conference: Optical Character Recognition in Mass Digitisation." *Ariadne,* 59 (June 2009). http://www.ariadne.ac.uk/issue59/impact-2009-rpt/

[Porter 2010]. Porter, Dorothy. "How Does TILE relate to TEI." *TILE Blog*, (March 13, 2010). http://mith.info/tile/2010/03/13/how-does-tile-relate-to-tei/

[Porter et al. 2006]. Porter, Dorothy, William Du Casse, Jerzy W. Jaromczyk, Neal Moore, Ross Scaife and Jack Mitchell. "Creating CTS Collections." *Digital Humanities 2006*, (2006): 269-74. http://www.csdl.tamu.edu/~furuta/courses/06c_689dh/dh06readings/DH06-269-274.pdf

[Porter et al. 2009]. Porter, Dorothy, Doug Reside, and John Walsh. "Text-Image Linking Environment (TILE)." *Digital Humanities Conference Abstracts 2009*, (June 2009): 388-390. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf

[Pritchard 2008]. Pritchard, David. "Working Papers, Open Access, and Cyber-Infrastructure in Classical Studies." *Literary & Linguistic Computing*, 23 (June 2008): 149-162. Preprint available at: http://ses.library.usyd.edu.au/handle/2123/2226

[Pybus and Kirkham 2009]. Pybus, John and Ruth Kirkham. "Experiences of User Involvement in the Construction of a Virtual Research Environment for the Humanities." *2009 5th IEEE International Conference on E-Science Workshops*, IEEE, (2009): 135-137. http://dx.doi.org/10.1109/ESCIW.2009.5407961

[Reddy and Crane 2006]. Reddy, Sravana and Gregory Crane. "A Document Recognition System for Early Modern Latin." *DHCS 2006-Chicago Colloquium on Digital Humanities and Computer Science*, (November 2006). http://hdl.handle.net/10427/57011

[Remondino et al. 2009]. Remondino, Fabio, Stefano Girardi, Alessandro Rizzi, and Lorenzo Gonzo. "3D Modeling of Complex and Detailed Cultural Heritage Using Multi-Resolution Data." *Journal of Computing and Cultural Heritage*, 2 (2009): 1-20. http://dx.doi.org/10.1145/1551676.1551678

[Reside 2010]. Reside, Doug. "A Four Layer Model for Image-Based Editions." *TILE Blog*, (February 2010). Part One: http://mith.info/tile/2010/02/03/a-four-layer-model-for-image-based-editions/ and Part Two: http://mith.info/tile/2010/02/

[Robertson 2009]. Robertson, Bruce. "Exploring Historical RDF with Heml." *Digital Humanities Quarterly*, 3 (2009). http://www.digitalhumanities.org/dhq/vol/003/1/000026.html

[Robinson 2009]. Robinson, Peter. "Towards a Scholarly Editing System for the Next Decades." *Sanskrit Computational Linguistics* (*Lecture Notes in Computer Science*, Volume 5402), Springer, (2009): 346-357. http://dx.doi.org/10.1007/978-3-642-00155-0_18

[Robinson 2010]. Robinson, Peter. "Editing Without Walls." *Literature Compass*, 7 (2010): 57-61. http://dx.doi.org/10.1111/j.1741-4113.2009.00676.x

[Rockwell 2010]. Rockwell, Geoffrey. "As Transparent as Infrastructure: On the Research of Cyberinfrastructure in the Humanities." *Online Humanities Scholarship: The Shape of Things to Come*, Rice University Press, (March 2010). http://rup.rice.edu/cnx_content/shape/m34315.html

[Romanello 2008]. Romanello, Matteo. "A Semantic Linking Framework to Provide Critical Value-Added Services for E-Journals on Classics." *ELPUB2008. Open Scholarship: Authority, Community, and Sustainability in the Age of Web 2.0 - Proceedings of the 12th International Conference on Electronic Publishing*, (June 2008): 401-414. http://elpub.scix.net/cgi-bin/works/Show?401_elpub2008

[Romanello et al. 2009a]. Romanello, Matteo, Federico Boschetti, and Gregory Crane. "Citations in the Digital Library of Classics: Extracting Canonical References By Using Conditional Random Fields." *NLPIR4DL '09: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*. Morristown, NJ, USA: Association for Computational Linguistics, (2009): 80-87. http://aye.comp.nus.edu.sg/nlpir4dl/NLPIR4DL10.pdf

[Romanello et al. 2009b]. Romanello, Matteo, Monica Berti, Federico Boschetti, Alison Babeu, and Gregory Crane. "Rethinking Critical Editions of Fragmentary Texts By Ontologies." *ELPUB2009. Proceedings of 13th International Conference on Electronic Publishing: Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies*, (2009): 155-174. Preprint available at: http://www.perseus.tufts.edu/publications/elpub2009.pdf

[Romanello et al. 2009c]. Romanello, Matteo, Monica Berti, Alison Babeu, and Gregory Crane. "When Printed Hypertexts Go Digital: Information Extraction from the Parsing of Indices." *HT '09: Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*. New York, NY, USA: ACM, (2009): 357-358. Preprint available at: http://www.perseus.tufts.edu/publications/ht159-romanello.pdf

[Romary and Armbruster 2009]. Romary, Laurent and Chris Armbruster. "Beyond Institutional Repositories." *Social Science Research Network Working Paper Series*, (June 2009). http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID1425919_code434782.pdf?abstractid=1425692&mirid=1

[Roueché 2009]. Roueché, Charlotte. "Digitizing Inscribed Texts." in *Text Editing, Print and the Digital World*. Ashgate Publishing, 2009, pg.159-169.

[Roued 2009] Roued, Henriette. "Textual Analysis Using XML: Understanding Ancient Textual Corpora." *5th IEEE Conference on e-Science 2009*, (December 2009). http://esad.classics.ox.ac.uk/index.php?option=com_docman&task=doc_download&gid=30&Itemid=78

[Roued-Cunliffe 2010]. Roued-Cunliffe, Henriette. "Towards a Decision Support System For Reading Ancient Documents." Forthcoming. *Literary and Linguistic Computing*, (2010). http://esad.classics.ox.ac.uk/index.php?option=com_docman&task=doc_download&gid=29&Itemid=97

[Ruddy 2009]. Ruddy, David. "Linking Resources in the Humanities: Using OpenURL to Cite Canonical Works." *DLF Spring 2009 Forum*, (May 2009). http://www.diglib.org/forums/spring2009/presentations/Ruddy.pdf

[Rudman 1998]. Rudman, Joseph. "Non-Traditional Authorship Attribution Studies in the *Historia Augusta*: Some Caveats." *Literary & Linguistic Computing*, 13 (September 1998): 151-157. http://dx.doi.org/10.1093/llc/13.3.151

[Ruhleder 1995]. Ruhleder, Karen. "Reconstructing Artifacts, Reconstructing Work: From Textual Edition to On-Line Databank." *Science, Technology, & Human Values*, 20 (1995): 39-64. http://www.jstor.org/stable/689880

[Rydberg-Cox 2002]. Rydberg-Cox, Jeffrey A. "Mining Data from an Electronic Greek Lexicon." *The Classical Journal*, 98 (2002): 183-188.

[Rydberg-Cox 2009]. Rydberg-Cox, Jeffrey A. "Digitizing Latin Incunabula: Challenges, Methods, and Possibilities." *Digital Humanities Quarterly*, 3 (January 2009). http://www.digitalhumanities.org/dhq/vol/3/1/000027.html#

[Salerno et al. 2007]. Salerno, Emanuele, Anna Tonazzini, and Luigi Bedini. "Digital Image Analysis to Enhance Underwritten Text in the Archimedes Palimpsest." *International Journal on Document Analysis and Recognition*, 9 (April 2007): 79-87. http://dx.doi.org/10.1007/s10032-006-0028-7

[Sankar et al. 2006]. Sankar, K., Vamshi Ambati, Lakshmi Pratha, and C. V. Jawahar. "Digitizing a Million Books: Challenges for Document Analysis." *Document Analysis Systems VII* (*Lecture Notes in Computer Science*, Volume 3872), (2006): 425-436. http://cvit.iiit.ac.in/papers/pramod06Digitizing.pdf

[Sayeed and Szpakowicz 2004]. Sayeed, Asad B. and Stan Szpakowicz. "Developing a Minimalist Parser for Free Word Order Languages with Discontinuous Constituency." *Advances in Natural Language Processing* (*Lecture Notes in Computer Science* (Volume 3230), (2004): 115-126.

[Schibel and Rydberg-Cox 2006]. Schibel, Wolfgang and Jeffrey A. Rydberg-Cox. "Early Modern Culture in a Comprehensive Digital Library." *D-Lib Magazine*, 12 (March 2006). http://www.dlib.org/dlib/march06/schibel/03schibel.html

[Schilit and Kolak 2008]. Schilit, Bill N. and Okan Kolak. "Exploring a Digital Library Through Key Ideas." *JCDL '08: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY, USA: ACM, 2008, 177-186. http://schilit.googlepages.com/fp035-schilit.pdf

[Schloen 2001]. Schloen, J. David. "Archaeological Data Models and Web Publication Using XML." *Computers and the Humanities*, 35 (May 2001): 123-152. http://dx.doi.org/10.1023/A:1002471112790

[Schmidt 2010]. Schmidt, Desmond. "The Inadequacy of Embedded Markup for Cultural Heritage Texts." *Literary and Linguistic Computing*, 25 (April 2010): 337-356. http://dx.doi.org/10.1093/llc/fqq007

[Schmidt and Colomb 2009]. Schmidt, Desmond and Robert Colomb. "A Data Structure for Representing Multi-Version Texts Online." *International Journal of Human Computer Studies*, 67 (June 2009): 497-514.
[Schoepflin 2003]. Schoepflin, Urs. "The Archimedes Project: Realizing the Vision of an Open Digital Research Library for the Study of Long-Term Developments in the History of Mechanics." *RCDL 2003*: *Proceedings of the 5th National Russian Research Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections,"* (2003): 124-129. http://edoc.mpg.de/get.epl?fid=15799&did=169534&ver=0

[Schonfeld and Housewright 2010]. Schonfeld, Roger C. and Ross Housewright. *Faculty Survey 2009: Key Strategic Insights for Libraries, Publishers, and Societies*. Ithaka, (April 2010). http://www.ithaka.org/ithaka-s-r/research/faculty-surveys-2000-2009/Faculty%20Study%202009.pdf

[Schmitz 2009]. Schmitz, Patrick. "Using Natural Language Processing and Social Network Analysis to Study Ancient Babylonian Society." *UC Berkeley iNews*, (March 1, 2009). http://inews.berkeley.edu/articles/Spring2009/BPS

[Schreibman 2009]. Schreibman, Susan. "An E-Framework for Scholarly Editions." *Digital Humanities 2009 Conference Abstracts*, (June 2009): 18-19. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf

[Schreibman et al. 2009]. Schreibman, Susan, Jennifer Edmond, Dot Porter, Shawn Day, and Dan Gourley. "The Digital Humanities Observatory: Building a National Collaboratory." *Digital Humanities 2009 Conference Abstracts*, (June 2009): 40-43. http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf

[Sennyey 2009]. Sennyey, Pongracz, Lyman Ross, and Caroline Mills. "Exploring the Future of Academic Libraries: A Definitional Approach." *The Journal of Academic Librarianship*, 35 (May 2009): 252-259. http://dx.doi.org/10.1016/j.acalib.2009.03.003

[Shen et al. 2008]. Shen, Rao, Naga Srinivas Vemuri, Weiguo Fan, and Edward A. Fox. "Integration of Complex Archeology Digital Libraries: An ETANA-DL Experience." *Information Systems*, 33 (November 2008): 699-723. http://dx.doi.org/10.1016/j.is.2008.02.006

[Shiaw et al. 2004]. Shiaw, Horn Y., Robert J. K. Jacob, and Gregory R. Crane. "The 3D Vase Museum: a New Approach to Context in a Digital Library." *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*. New York, NY, USA: ACM, (2004): 125-134. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=0AAF0CD5C3D796C85A4E7EC96EA8FDCD?doi=10.1.1.58.760&rep=rep1&type=pdf

[Shilton 2009]. Shilton, Katie. *Supporting Digital Tools for Humanists: Investigating Tool Infrastructure*. *Final Report*. Council for Library and Information Resources, (May 2009). http://www.clir.org/pubs/archives/ShiltonToolsfinal.pdf

[Siemens 2009]. Siemens, Lynne. "'It's a Team if You Use "Reply All"': An Exploration of Research Teams in Digital Humanities Environments." *Literary and Linguistic Computing*, 24 (April 2009): 225-233.

[Smith 2002]. Smith, David A. "Detecting Events with Date and Place Information in Unstructured Text." *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY, USA: ACM, (2002): 191-196. Preprint available at: http://www.perseus.tufts.edu/publications/datestat.pdf

[Smith 2008]. Smith, Abby. "The Research Library in the 21st Century: Collecting, Preserving, and Making Accessible Resources for Scholarship." in *No Brief Candle: Reconceiving Research Libraries for the 21st Century*. Council on Library and Information Resources, Publication Number 142, (August 2008): pp. 13-20. http://www.clir.org/pubs/reports/pub142/pub142.pdf

[Smith 2009]. Smith, D. Neel. "Citation in Classical Studies." *Digital Humanities Quarterly*, 3 (2009). http://www.digitalhumanities.org/dhq/vol/003/1/000028.html#

[Smith 2010]. Smith, D. Neel. "Digital Infrastructure and the Homer Multitext Project." in *Digital Research in the Study of Classical Antiquity* (eds. Gabriel Bodard and Simon Mahony). Burlington, VT: Ashgate Publishing, 2010, pp.121-137.

[Snow et al. 2006]. Snow, Dean R., Mark Gahegan, Lee C. Giles, Kenneth G. Hirth, George R. Milner, Prasenjit Mitra, and James Z. Wang. "Cybertools and Archaeology." *Science*, 311 (February 2006): 958-959. http://dx.doi.org/10.1126/science.1121556

[Speck 2005]. Speck, Reto. *The AHDS and Digital Resource Creation in Classics, Ancient History, Philosophy, Religious Studies and Theology*. Arts and Humanities Data Service, (November 2005). http://ahds.ac.uk/about/projects/documents/subject_extension_report_v1.pdf

[Stewart et al. 2007]. Stewart, Gordon, Gregory Crane, and Alison Babeu. "A New Generation of Textual Corpora: Mining Corpora from Very Large Collections." *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint Conference on Digital Libraries*. New York, NY, USA: ACM, (2007): 356-365.
Preprint available at: http://hdl.handle.net/10427/14853

[Stokes 2009]. Stokes, Peter A. "Computer-Aided Palaeography, Present and Future." *Digital Humanities 2009 Conference Abstracts*, (June 2009): 266-268.
http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf

[Stinson 2009]. Stinson, Timothy. "Codicological Descriptions in the Digital Age." *Kodikologie und Paläographie im digitalen Zeitalter-Codicology and Palaeography in the Digital Age.* Norderstedt: Books on Demand, 2009, pp. 35-51. Also available online at: http://kups.ub.uni-koeln.de/volltexte/2009/2959/

[Tablan et al 2006]. Tablan, Valentin, Wim Peters, Diana Maynard, and Hamish Cunningham. "Creating Tools for Morphological Analysis of Sumerian." *Proceedings of LREC 2006*, (2006).
http://gate.ac.uk/sale/lrec2006/etcsl/etcsl-paper.pdf

[Tambouratzis 2008]. Tambouratzis, George. "Using an Ant Colony Metaheuristic to Optimize Automatic Word Segmentation for Ancient Greek." *IEEE Transactions on Evolutionary Computation*, 13 (2009): 742-753.
http://dx.doi.org/10.1109/TEVC.2009.2014363

[Tarrant et al. 2009]. Tarrant, David, Ben O'Steen, Tim Brody, Steve Hitchcock, Neil Jefferies, and Leslie Carr. "Using OAI-ORE to Transform Digital Repositories into Interoperable Storage and Services Applications." *The Code4Lib Journal*, (March 2009). http://journal.code4lib.org/articles/1062

[Tarte 2010]. Tarte, Segolene M. "Papyrological Investigations: Transferring Perception and Interpretation into the Digital World." *Literary and Linguistics Computing*, (Forthcoming).
http://esad.classics.ox.ac.uk/index.php?option=com_docman&task=doc_details&gid=21&Itemid=97

[Tarte et al. 2009]. Tarte, Segolene M., David Wallom, Pin Hu, Kang Tang, and Tiejun Ma. "An Image Processing Portal and Web-Service for the Study of Ancient Documents." *5th IEEE conference on e-Science 2009*, (December 2009): 14-19.
http://esad.classics.ox.ac.uk/index.php?option=com_docman&task=doc_download&gid=20&Itemid=78

[Tchernetska et al. 2007]. Tchernetska, Natalie, E. Handley, C. Austin, L. Horváth, "New Readings in the Fragment of Hyperides' Against Timandros from the Archimedes Palimpsest." *Zeitschrift für Papyrologie und Epigraphik*, 162 (2007): 1-4.

[Terras 2005]. Terras, Melissa. "Reading the Readers: Modelling Complex Humanities Processes to Build Cognitive Systems." *Literary and Linguistic Computing*, 20 (March 2005): 41-59.
http://dx.doi.org/10.1093/llc/fqh042

[Terras 2010]. Terras, Melissa. "The Digital Classicist: Disciplinary Focus and Interdisciplinary Vision." In *Digital Research in the Study of Classical Antiquity* (eds. Gabriel Bodard and Simon Mahony). Burlington, VT: Ashgate Publishing, 2010, pp. 171-189.

[Thiruvathukal 2009]. Thiruvathukal, George K., Steven E. Jones, and Peter Shillingsburg. "The e-Carrel: An Environment for Collaborative Textual Scholarship." *DHCS 2009,* (October 2009). http://lingcog.iit.edu/~argamon/DHCS09-Abstracts/Thiruvathukal.pdf

[Tobin et al. 2008]. Tobin, Richard, Claire Grover, Sharon Givon, and Julian Ball. "Named Entity Recognition for Digitised Historical Texts." *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08),* (2008). http://www.ltg.ed.ac.uk/np/publications/ltg/papers/bopcris-lrec.pdf

[Toms and Flora 2005]. Toms, Elaine G. and N. Flora. "From Physical to Digital Humanities Library – Designing the Humanities Scholar's Workbench." in *Mind Technologies: Humanities Computing and the Canadian Academic Community* (eds. R. Siemens and D. Moorman). Calgary, Canada: University of Calgary Press, 2005

[Toms and O'Brien 2008]. Toms, Elaine G. and Heather L. O'Brien. "Understanding the Information and Communication Technology Needs of the E-Humanist." *Journal of Documentation*, 64 (2008): 102-130.

[Tonkin 2008]. Tonkin, Emma. "Persistent Identifiers: Considering the Options." *Ariadne,* 56 (July 2008), http://www.ariadne.ac.uk/issue56/tonkin/

[Toth and Emery 2008]. Toth, Michael and Doug Emery. "Applying DCMI Elements to Digital Images and Text in the Archimedes Palimpsest Program." *DC-2008—Berlin Proceedings of the International Conference on Dublin Core and Metadata Applications,* (2008): 163-168. http://dcpapers.dublincore.org/ojs/pubs/article/view/929

[Toufexis 2010]. Toufexis, Notis. "One Era's Nonsense, Another's Norm: Diachronic Study of Greek and the Computer." in *Digital Research in the Study of Classical Antiquity* (eds. Gabriel Bodard and Simon Mahony). Burlington, VT: Ashgate Publishing, 2010, pp. 105-118.

[Tupman 2010]. Tupman, Charlotte. "Contextual Epigraphy and XML: Digital Publication and its Application to the Study of Inscribed Funerary Monuments." in *Digital Research in the Study of Classical Antiquity* (eds. Gabriel Bodard and Simon Mahony). Burlington, VT: Ashgate Publishing, 2010, pp. 73-86.

[Unsworth 2000]. Unsworth, John. "Scholarly Primitives: What Methods do Humanities Researchers Have in Common, and How Might our Tools Reflect This?" (2000). http://www.iath.virginia.edu/~jmu2m/Kings.5-00/primitives.html

[Uytvanck et al. 2010]. Uytvanck, Dieter V., Claus Zinn, Daan Broeder, Peter Wittenburg, and Mariano Gardellini. "Virtual Language Observatory: The Portal to the Language Resources and Technology Universe." *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).* European Language Resources Association (ELRA), (May 2010). http://www.lrec-conf.org/proceedings/lrec2010/pdf/273_Paper.pdf

[van Groningen 1932]. van Groningen, B. A. "Projet d'unification des Systemes de Signes Critiques." *Chronique d'Egypte,* 7 (13-14), (1932): 262-269.

[van de Sompel and Lagoze 2007]. van de Sompel, Herbert and Carl Lagoze. "Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication." *CTWatch Quarterly,* 3 (August 2007). http://www.ctwatch.org/quarterly/articles/2007/08/interoperability-for-the-discovery-use-and-re-use-of-units-of-scholarly-communication/

[Váradi et al. 2008]. Váradi, Tamás, Steven Krauwer, Peter Wittenburg, Martin Wynne, and Kimmo Koskenniemi. "CLARIN: Common Language Resources and Technology Infrastructure." *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), (May 2008). http://www.lrec-conf.org/proceedings/lrec2008/summaries/317.html

[Vemuri et al. 2006]. Vemuri, Naga S., Rao Shen, Sameer Tupe, Weiguo Fan, and Edward A. Fox. "Etana-ADD: An Interactive Tool For Integrating Archaeological DL Collections." *JCDL '06: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY, USA: ACM Press, (2006): 161-162.

[Vertan 2009]. Vertan, Cristina. "A Knowledge Web-Based eResearch Environment for Classical Philology." *Digital Classicist-Works in Progress Seminar,* (July 2009). http://www.digitalclassicist.org/wip/wip2009-06cv.pdf

[Villegas and Parra 2009] Villegas, Marta and Carla Parra. "Integrating Full-Text Search and Linguistic Analyses on Disperse Data for Humanities and Social Sciences Research Projects." *Fifth IEEE International Conference on E-Science*, (December 2009): 28-32. http://dx.doi.org/10.1109/e-Science.2009.12

[Vlachopoulos 2009]. Vlachopoulos, Dimitrios. "Introducing Online Teaching in Humanities: A Case Study About the Acceptance of Online Activities by the Academic Staff of Classical Languages." *DIGITHUM*, 11 (May 2009). http://digithum.uoc.edu/ojs/index.php/digithum/article/view/n11_vlachopoulos

[Voss and Procter 2009]. Voss, Alexander and Rob Procter. "Virtual Research Environments in Scholarly Work and Communications." *Library Hi Tech*, 27 (2009): 174-190. http://dx.doi.org/10.1108/07378830910968146

[Vuillemot 2009]. Vuillemot, Romain, Tanya Clement, Catherine Plaisant, and Amit Kumar. "What's Being Said Near "Martha"? Exploring Name Entities in Literary Text Collections." *VAST 2009: IEEE Symposium on Visual Analytics Science and Technology*, (2009): 107-114. http://dx.doi.org/10.1109/VAST.2009.5333248

[Wallom et al. 2009]. Wallom, David, Segolene Tarte, Tiejun Ma, Pin Hu, and Kang Tang. "Integrating eSAD (The Image, Text, Interpretation: e-Science, Technology and Documents) and VRE-SDM (VRE for the Study of Documents and Manuscripts) projects." *AHM 2009*, (2009). http://esad.classics.ox.ac.uk/index.php?option=com_docman&task=doc_download&gid=23&Itemid=78

[Warwick et al. 2008a]. Warwick, Claire, Melissa Terras, Paul Huntington, and Nikoleta Pappa. "If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities Through Statistical Analysis of User Log Data." *Literary & Linguistic Computing*, 23 (April 2008): 85-102.

[Warwick et al. 2009]. Warwick, Claire, Claire Fisher, Melissa Terras, Mark Baker, Amanda Clarke, Mike Fulford, Matt Grove, Emma O'Riordan, and Mike Rains. "iTrench: A Study of User Reactions to the Use of Information Technology in Field Archaeology." *Literary & Linguistic Computing*, 24 (June 2009): 211-223.

[Warwick et al. 2008b]. Warwick, Claire, Isabel Galina, Melissa Terras, Paul Huntington, and Nikoleta Pappa. "The Master Builders: LAIRAH Research on Good Practice in the Construction of Digital Humanities Projects." *Literary & Linguistic Computing*, 23 (September 2008): 383-396.

[Weenink et al. 2008]. Weenink, Kasja, Leo Waaijers, and Karen van Godtsenhoven. *A DRIVER's Guide to European Repositories: Five Studies of Important Digital Repository Related Issues and Good Practices*. Amsterdam University Press, (2008). http://dare.uva.nl/document/93898

[Wynne et al. 2009]. Wynne, Martin, Steven Krauwer, Sheila Anderson, Chad Kainz, and Neil Fraistat. "Supporting the Digital Humanities: Putting the Jigsaw Together." *Digital Humanities Conference Abstracts 2009*, (June 2009): 49.
http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf

[Zielinski et al. 2009]. Zielinski, Andrea, Wolfgang Pempe, Peter Gietz, Martin Haase, Stefan Funk, and Christian Simon. "TEI Documents in the Grid." *Literary & Linguistic Computing*, 24 (September 2009): 267-279.

[Zorich 2008]. Zorich, Diane M. *A Survey of Digital Humanities Centers in the United States*. Council on Library and Information Resources, Publication Number 143, (April 2008).
http://www.clir.org/pubs/abstract/pub143abst.html