DHQ: Digital Humanities Quarterly
Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure
2009
Volume 3 Number 1

Conclusion: Cyberinfrastructure, the Scaife Digital Library and Classics in a Digital age

Christopher Blackwell <Christopher_dot_Blackwell_at_furman_dot_edu>, Furman University
Gregory Crane <gregory_dot_crane_at_tufts_dot_edu>, Tufts University

Abstract

We can already begin to envision research projects that were scarcely, if at all, feasible in print culture. The papers in this collection allow us as well to enumerate the services and publication types on which emerging scholarship depends. We also need models for publication that meet the needs and realize the potential of the digital media and we describe here the Scaife Digital Library, a concrete example of true digital publication.

> I look upon the discontent of the literary class, as a mere announcement of the fact, that they find themselves not in the state of mind of their fathers, and regret the coming state of mind as untried, as a boy dreads the water before he has learned that he can swim. If there is any period one would desire to be born in, is it not the age of Revolution; when the old and the new stand side by side, and admit of being compared; when the energies of all men are searched by fear and by hope; when the historic glories of the old, can be compensated by the rich possibilities of the new era? This time, like all times, is a very good one, if we but know what to do with it. (Emerson, *The American Scholar*)

> Every human individuality is an idea rooted in actuality, and this idea shines forth so brilliantly from some individuals that it seems to have assumed the form of an individual merely to use it as a vehicle for expressing itself. When one traces human activity, after all its determining causes have been subtracted there remains something original which transforms these influences instead of being suffocated by them; in this very element there is an incessantly active drive to give outward shape to its inner, unique nature. (Wilhelm von Humboldt, “Lecture to the Prussian Academy,” 1821)

When Emerson addressed Harvard’s Phi Beta Kappa Society in 1837, slavery was still an established institution and those who in Massachusetts favored its abolition, such as William Lloyd Garrison, were the dangerous radicals of their day and those who, like the author Lydia Maria Child, suggested racial equality found the doors of polite society slamming shut in their faces. Many twenty-first century readers will note the linguistic assumption that scholars are boys, fathers and men. Revolution has its own logic and revolutionaries should never forget that the critical pose which they apply to the present and the past will turn itself upon them when they have themselves passed into history — if, of course, they are so fortunate as to touch the historical memory of succeeding generations. If, in decades and generations to come, students of the ancient world read these words, we cannot now say where they may pause to wonder at how prescient the members of this early generation had been or where they may cringe and squirm. But all of those who contributed to this collection have dedicated their lives to a love for the past and that love allows us to embrace the future. The authors of this collection cannot predict what course events will assume or how they will appear to those who follow, but they have recognized the revolution of their own time and all have taken action to carry this revolution forward. 1

Emerson does not really define the title of his talk, but for those of us who contributed to this collection, whether we happen to live in United States or not, Ross Scaife embodied the best qualities that a phrase such as the “American 2

Scholar" might suggest. Of course, Ross happens to have lived his life in the United States: born and educated in Virginia, trained as a scholar in Texas, Ross fashioned a home in Kentucky — and the many who had the privilege of visiting that home know how much literal truth there is in that statement. But Ross was a man unmoved by social convention or established authority. For him, the future was one of boundless possibility and the past was not a burden, but a foundation on which to build. He feared neither change nor continuity but evaluated each on its own merits according to the values that had grown strong within his heart. And like every true scholar from every nation and every period, he loved both grand ideas and the people around him with equal warmth.

A generation from now, the course that classical studies and the humanities in general have taken may seem to have been a natural outgrowth of the early twenty-first century. And, indeed, we cannot say to what extent the larger forces at work within society may constrain the shape that our field will assume. But those of us who knew Ross also saw a man who anticipated far ahead of his fellows the importance of making our ideas accessible to the widest possible audience. The original proposal that secured funding to the Stoa called for a new generation of publications that would be designed from the start to be intellectually as well as physically accessible to an audience far beyond the narrow channels of twentieth century academic discourse. Blackwell and Martin in this collection articulate how this vision was, in fact, realized: Stoa publications such as Blackwell's *Demos* [1] bring the broader public directly into contact both with his interpretation of Athenian democracies and with the primary sources on which his interpretations are based. 3

Ross was among the first to recognize the importance of making our publications fully open — it is not enough to provide a single perspective via a single web site with primary and secondary sources. We need to make the source materials accessible — others need to be able to download what we produce, apply their own analytical methods, and even build new derivative works on what others have done. It is already difficult for us to remember how radical and far-sighted Ross was years ago. He had the vision to see what was obviously wrong at the time but would become obviously correct in the future. Ross embodied that profound originality that Humboldt describes in those who produce the times of which we are all products. 4

In this conclusion, we synthesize some of the themes outlined and work described in the previous papers. We recall the categories of ePhilology and eClassics, first discussed in the introduction, and use these two categories to characterize two fundamental advances now becoming possible: our ability to begin increasingly complex intellectual projects with greater command of the underlying data and to answer finally the challenge, articulated in Plato's *Phaedrus*, that written words cannot explain themselves. We then shift to describe some of the basic services and collections that must be a part of any Cyberinfrastructure for classics and humanities. From there, we list the requirements for publication for a Cyberinfrastructure in which automated systems and broad based communities interact in novel, complex ways with our primary and secondary sources. We then describe the Scaife Digital Library (SDL), an open effort that integrates primary and secondary sources, has an immediate core of classical materials but also can manage content from many disciplines, and embodies more broadly and perfectly than any other effort with which we are familiar the needs of advanced research. While the SDL represents, in our view, a major step forward for classical studies and ultimately, we hope, for other disciplines, the SDL builds directly upon foundations that Ross Scaife laid over his decade of work on the Stoa Publishing Consortium.[2] 5

Opportunities: ePhilology and eClassics

> And one day they taught Hesiod glorious song while he was shepherding his lambs under holy Helicon, and this word first the goddesses said to me — the Muses of Olympus, daughters of Zeus who holds the aegis:
>
> "Shepherds of the wilderness, wretched things of shame, mere bellies, we know how to speak many false things as though they were true; but we know, when we will, to utter true things."
>
> So said the ready-voiced daughters of great Zeus, and they plucked and gave me a staff, a shoot of sturdy laurel, a marvelous thing, and breathed into me a divine voice to celebrate things that shall be and things that were before; and they bade me sing of the race of the blessed gods that are eternally, but ever

> to sing of themselves both first and last. (Hesiod, *Theogony* 21-34, after Evelyn-White)

The Muses gave Hesiod a staff, and for the poet that is enough — few, if any, have produced poetry that has exerted such a spell over so many people from so many periods of time and disparate cultures as have the works of Hesiod and the Homeric Epics. All of us who live the life of the mind, whether we are poets or professors, follow our Muses. The staff that we have now taken into our hands is still rough and we are learning its balance and heft, but already we can begin to glimpse the stories that we will be able to see when the inspiration of our new muses takes full hold. 6

The introduction to this collection distinguished two goals within a digital world. On the one hand, ePhilology emphasizes the role of the linguistic record in producing and organizing ideas and information about the ancient world. We use eClassics, by contrast, to describe Greek and Latin languages and literatures, wherever and whenever produced, as they live within our physical brains, touch our less tangible hearts and shape our actions in the world around us. We return now to these topics, suggesting how a Cyberinfrastructure, including both comprehensive collections and advanced, domain optimized services, can advance each of these goals. Memographies allow philologists to explore vast topics far too large for individual scholars in print culture. Plato’s challenge allows us to appreciate the magnitude of the opportunities before us now, as we can finally begin to address a critique of the static written word that is more than two thousand years old. 7

ePhilology and Memographies

> My mother Thetis tells me that there are two ways in which I may meet my end. If I stay here and fight, I shall lose my safe homecoming but I will have a glory that is unwilting: whereas if I go home my glory will die, but it will be a long time before the outcome of death shall take me. (Achilles’ choice, Homer, *Iliad* 9.410-416, tr. Butler/Nagy)

It is easy to see how we can, in a digital environment, pursue our research topics more extensively than was previously possible. We have also described how we can make the sources of antiquity intellectually accessible to new audiences. We now turn to the question of what research questions we can pursue that would not have been feasible without collections that are, if not exhaustive, at least large enough to be representative of the published record available in print. 8

Consider a monolingual printed corpus such as English language newspapers in the 19th century United States. The 1869 *Rowell Newspaper Directory* [3] for the United States and Canada lists more than 5000 newspapers that were printing more than 20,000 unique pages a week and thus more than 1,000,000 pages per year. If we take 5,200 pages of one newspaper (the Civil War era *Richmond Times Dispatch* [4]) as a rough indicator of words on a typical newspaper page (c. 5,000), North American English language newspapers printed perhaps 50 billion words each year in the late 1860s. If we simply analyzed these newspapers, we could open up whole new lines of inquiry, tracking a range of topics: Which newspapers reprinted stories from which? What sorts of things did people say in newspapers from different parts of the country with different party affiliations about slavery over the course of time? What poetry and fiction appeared in these newspapers? What products were advertised? All of these are eminently tractable problems: we don’t need perfect transcriptions or perfect services to begin identifying the trends behind these topics. If we begin to think about 19th century newspapers in other languages around the world, the challenges and opportunities become even greater. 9

Clearly we can begin to pursue topics that require analysis of much more data than any human being can see, much less contemplate. We can begin to trace topics that have a life in human tradition that goes beyond any single period or immediate context. Such topics have lives of their own. We can now write histories or (to pursue the metaphor of living things) biographies of these topics. The geneticist Richard Dawkins coined the term *meme* in 1976 to describe the cultural counterpart to biological genes: memes include any thoughts or behaviors that can be passed from one person to another and examples include “thoughts, ideas, theories, gestures, practices, fashions, habits, songs and dances.” [5] The term *meme* provides a useful concept because it stresses the autonomy of ideas as they circulate through our biological brains and storage technologies. The concept of a meme allows us to consider both information about a 10 historical topic that existed in the material world (e.g., the life of the historical Alexander the Great) and topics that have a life of their own (e.g., Alexander as a hero of Iranian folk tales). We use the term memography to describe the history of a meme within a larger body of material.

11 While the term *meme* may be new, the underlying concept is not. The continuous tradition of European literature begins with the *Iliad* and Achilles' choice. He knows from his mother, the goddess Thetis that he may choose a long but unremarkable life, soon forgotten, or he may die young but win undying fame. Undying fame means that others will speak about him and what he accomplished at Troy forever. He can trade life as a biological entity for life as an idea that the songs of the epic tradition will pass from one mind to another. The physical Achilles will pass sooner or later. But this new entity — this object of thought and memory — will never die. Achilles knows ahead of time that his death would secure for him the goal of all great heroes: he would become a *meme* — a meme that succeeded because it jumped from the medium of oral transmission into a network of material information technologies.

12 The biological Plato, likewise, vanished more than two thousand years ago but his writings have been copied ever since and the historical Plato continues to exist as the topic of discourse. Scholars could, in print culture before the advent of searchable texts, laboriously track down many Platonic testimonia, e.g., the explicit quotations and most obvious allusions to particular passages in Plato. German classicists have begun to apply text mining algorithms to search for quotations and allusions that previous generations missed.[6] If we wanted to understand the role of Plato and the ways in which others have quoted and used his dialogues, we would need to work in every language where Plato was influential. This would include not only such common languages of classical philology as Latin, English, French, German and Italian, but virtually every European language that left behind a substantial body of written discourse. If we then consider that Plato has had a major presence within Islamic thought and realize that we will need to consider Arabic and Persian as well, it quickly becomes clear that no single scholar can create from the primary sources a global overview of Plato's influence from antiquity through the present. The nineteenth-century newspapers mentioned above present just another component from the sources that shed light on who said what about Plato.

13 In an age of very large collections, we can, however, begin to design systems that will provide automatic visualizations of topics such as Plato and Plato's works.

- Named entity analysis finds passages that refer to Plato the philosopher, filtering out those passages that refer to other figures of the same name (e.g., the Athenian Comic poet named Plato).
- Quotation identification finds direct quotations and paraphrases of passages in Plato.
- Cross language information retrieval extends named entity and quotation identification to multiple languages (e.g., Arabic, Chinese, Latin, English, French, German, Italian, Russian and other languages for which major cross-lingual resources are available).
- Text mining identifies words and phrases that appear in conjunction with references to and quotations of Plato. These words and phrases allow us to discover common ideas associated with Plato across different genres and periods.
- Machine translation links similar words and phrases associated with Plato in multiple languages, identifying cross-lingual cultural units.
- Visualization systems allow readers to track, for example, where and how often Plato's Republic has been discussed, what passages have been most examined, and what sorts of things people have said about Plato, whether in Berlin or the Iranian university city of Qom.
- Customization and personalization services then provide individual analysts with relevant materials in languages that they understand as well as machine translation and interactive translation support services to help them with languages in which they have little or no fluency. Thus, the system might present scholars of Islamic thought with translations of Plato and translation support geared to their particular knowledge of Greek.

14 Each of the above and similar processes is analogous to the sensors by which scientists track data in the material world. Each of the above processes will produce noise as well as a usable signal. The results will not, of course, be scholarship, but rather data within which patterns can emerge to stimulate scholarship — in the end, human beings will

have to contemplate what the systems have found. They will refine the questions that they ask, contemplate the results again, and then repeat their analysis in an iterative process. But, despite all the noise within the system, we will quickly start to see patterns about who has said what at various times about which passages of Plato in a variety of languages.

If we consider established genres of reference work such as lexica, grammars, encyclopedias and editions, we can see that a wide range of topics constitute memes that we could now begin to study. 15

- People and places: Any major person (Shakespeare, Abraham Lincoln) and place (Rome, Athens) has a history within human imagination that goes far beyond anything that we could analyze with traditional means.
- Languages: Few scholars study Latin much as they may wish to: no one can become sufficiently familiar with all the communities who wrote in Latin for more than 2,000 years to describe the language as a whole. We can, however, already begin to track patterns of syntax, style, and lexicography as they change in different genres and periods.
- Abstract concepts: Some concepts aggressively attempt to transcend language and cultural barriers. Thus official Catholic doctrines are in theory designed to be comprehensible to any speaker of any language. The Pythagorean theorem both points to a mathematical concept and comprises a metaphor for mathematical knowledge with its own history.
- Texts: The Greek texts of Plato’s *Republic* and the Christian *New Testament* are both textual entities that have their own existence, in an open ended set of language versions, as complete versions, as the source for quotation, and as a foundation for allusion.

No one will ever be able to see, much less read and contemplate over time, the primary sources underlying broad topics such as the history of Latin over two thousand years or even the reception of Plato. Of course, this is hardly new: no living humanist publishing on major canonical authors such as Homer or Shakespeare can claim to have read and pondered more than a subset of conventional published scholarship in the conventional languages of European and American scholarship. But the rise of large collections and emergent systems with which to analyze those collections allows us to shift our stance away from the limits of what we can read with our two eyes and towards the challenges of working with machines that can scan large bodies of material and then (as we will see through the discussion of Plato’s challenge below) allow us to focus in detail on passages in more languages and from more contexts than was possible before.

A memography contains elements that are deeply traditional in form and general purpose, even if it represents an engagement between author, reader and source materials so quantitatively broader in scope as to constitute a radical change. *Demos* has only begun to adapt the scholarly monograph to a digital form but it already illustrates the increased connection between argument and primary source that we expect from a memography: *Demos* covers a major topic — Athenian Democracy — which had grown so heavily worked that many publications on the topic stopped providing direct citations to the primary sources on which their conclusions were based. *Demos* provides, wherever possible, not only citations to the primary sources on which each statement is based but also explanatory information — briefing materials, in effect — to support critical analysis of the sources once found. The history of Athenian democracy is a tractable subject. The history of the reception of Athenian democracy that includes Islamic and Western views of Athenian democracy is a memography. Thomas Martin’s *Overview of Greek Civilization* (also published separately as a book by Yale University Press)[7] was published as an on-line product for the original Perseus CD ROM publications and constitutes a subject so vast as to justify a memography. 16

A memography, in effect, applies the same principles to even larger topics and immediately requires automated methods. *Demos* exploited the digital environment of the early 21st century to more fully realize the ancient goals of the monograph. When Thucydides invented the form of the scholarly monograph, there were neither libraries of stable written sources nor, if such collections had existed, were there the systems whereby he could reliable cite these sources. In the historic passage where he describes his methodology, Thucydides reports that he has sifted the evidence and generated from this his best analysis of what happened (Thuc. 1.22). During the course of antiquity we find authors who begin to quote other authors and we begin to find references for previous works by author, work and 17 even chapter length “books” (the amount that a papyrus scroll could conveniently store). Print technology allowed us to refine these citations so that we could describe precise variations between multiple editions of a single work. *Demos* set out in the twenty-first century to restore to discussion on Athenian Democracy the connection between statements and primary sources that the citation system on which Roman historian Edward Gibbon could already in the eighteenth century rely.

Characteristics of a memography include: 18

- **Citation**: A memography contains citations between statements and the evidence on which they are based. A memography differs from a traditional monograph because in a memography we know that authors have only been able to scrutinize a subset of the evidence cited. Citations in a memography include versioned queries: we can thus see what evidence was available at the time when the memography was completed and how that evidence has subsequently changed as new sources come on-line, existing analytical tools become more powerful or wholly new services emerge.
- **Scale**: A project becomes a memography as its scope brings in more primary materials than a single human author can effectively analyze. Topics so vast that authors in print culture needed to focus their work on synthesizing specialized studies and could base their work primarily upon the primary sources would be subjects for memographies. The author must depend upon techniques such as sampling and automated analyses. A memography of George Washington would, for example, require, as one foundational dataset, the relative frequency of references to George Washington in multiple periods, genres, languages and cultural contexts. Such figures would require automated named entity analysis applied to very large collections. The memography would include a human author’s assessment of the accuracy of the automatically generated data.
- **Heterogeneity**: Memographies include not only more content than authors can review but content that assumes more categories of background knowledge than individual authors can expect to acquire. Such barriers can be language, cultural background, mathematics and any other topic. The history of mechanics could thus justify a memography because it requires not only a substantial understanding of mathematics and physics but sources produced over millennia and across Europe, North Africa and the Middle East in Greek, Latin, Arabic and every European language. Memographies thus require scalable, automated systems that can provide customized background information with which readers can examine and manually analyze any given object referenced. Thus, readers without training in Arabic but familiar with other languages and with the underlying scientific contexts can use automated morphological analyses, links to an on-line dictionary, and existing translations in languages that they do understand to pull apart Arabic source texts and determine which words are used in particular contexts to describe key concepts.[8]

Whether we are producing or reading (or both), most memographies will force us to interrogate primary materials from more contexts, linguistic, cultural or both, than we can expect to have studied in detail — the most powerful memes will work their way across time, genre, language and culture and it is this very quality that leaves a trail too long and complex for any single human mind. We must look to machines which can find and preprocess material relevant to a given meme through immense bodies of data.

19 The heterogeneity of background knowledge brings us again to the need for a Cyberinfrastructure. The German-US Archimedes Project was able to assemble the machine readable dictionaries, on-line source texts, morphological analyzers, annotation systems and other resources needed to explore the history of mechanics. Scholars without training in Arabic were, for example, able to work effectively with materials in Arabic. Almost two decades ago, a formal evaluation of students using the first generation of Perseus reading tools had already demonstrated that students with no knowledge of Greek could produce analyses of Greek texts that, in the view of external evaluators, matched the performance of students with advanced training in the language.[9] One major purpose of a Cyberinfrastructure is to generalize these results, providing a platform in which an increasing number of topics receive increasingly sophisticated services with even more dramatic results than those obtained by Perseus, Archimedes and other individual projects.

20 New technologies can help us locate relevant currents in the vast oceans of source material but we will still need to

descend from our overview and think carefully about some subset of the sources. While we will never be able to read everything, it becomes all the more important for us to ponder a few things, carefully selected, in great detail. In the past, practical issues such as language were fundamental barriers: if we found a text in a language that we could not read and did not have a human informant or translation, then we could literally do nothing. That condition has begun to change. This leads us to the topic of eClassics and Plato’s Challenge.

eClassics and Plato’s Challenge

Socrates:	Writing, Phaedrus, has this strange quality, and is very like painting; for the creatures of painting stand like living beings, but if one asks them a question, they preserve a solemn silence. And so it is with written words; you might think they spoke as if they had intelligence, but if you question them, wishing to know about their sayings, they always say only one and the same thing.
Socrates:	When one says “iron” or “silver” we all understand the same thing, do we not?
Phaedrus:	Surely.
Socrates:	What if he says “justice” or “goodness”? Do we not part company, and disagree with each other and with ourselves?
Phaedrus:	Certainly.

21 In a famous paper, published in 1950, Alan Turing proposed what has been since called the Turing test: a machine demonstrates intelligence when we cannot tell whether we are conversing with a human or a machine.[10] We propose a simpler challenge based upon a critique posed in Plato’s *Phaedrus*. In that dialogue, Plato’s Socrates critiques writing as inert and voiceless — we can no more ask the written word to explain itself than we can carry on a conversation with a painting, however lifelike it may appear. In a digital world, however, we can begin to address this ancient critique: manually edited hyperlinks and search engines are only initial instruments by which readers can make digital texts less opaque than their print counterparts. Named entity analysis addresses questions such as “to which *Alexander* does this particular passage refer?” and enables services such as plotting the right Alexandria for a given passage on a map for the relevant chronological period. Simple dictionary look-up tools answer questions such as “what does this word mean?” Word sense disambiguation systems allow us to determine the probability of a particular word sense in a given context (e.g., Latin *oratio* as “speech” vs. “prayer”). Text mining systems elicit key words and phrases by which documents can begin to describe what they are about. We may be a long way from a meaningful answer to the Turing test, but even relatively simple technologies have allowed us to make progress against the challenge that Plato leveled against information technology two and a half millennia ago.

22 Addressing Plato’s challenge has important implications for the problems that humanists choose to address. In 1972, Jacques Derrida published an essay, translated into English in 1981 as “Plato’s Pharmacy”,[11] that featured the critique of writing in the *Phaedrus*. In that dialogue, one speaker rejects the claim that writing aids memory — writing is not a medicine but a poison, encouraging us to depend upon writing and weakening our memories. Derrida’s essay probes the limitations of what we can express in language and thus, innovative as it may have seemed, reinforces the traditional scholarly focus upon questions that are obscure and may indeed have no final answer (e.g., topics prominent in the *Phaedrus* such as love and truth).

23 If we address the *Phaedrus* test, however, we find ourselves looking at scholarship from the opposite direction. Derrida pondered the limitations of language and logic. More traditional scholars such as the Latinist D. R. Shackleton Bailey pondered the best variant reading in a given text or to which Antonius a particular passage referred. Both focused upon the extremes, where language or the historical evidence at our disposal had not been sufficient for human analysis to

generate a final decision. Scholarship largely focused on outliers.

In addressing Plato’s challenge, we focus less upon the 2% of instances where we cannot readily determine to which Antonius an author refers than upon the other 98% where any reader, familiar with the context, can determine the intended referent. To address Plato’s challenge, we need to maximize a machine’s ability to recognize the dizzying number of simple referents that expert readers understand without conscious effort. We shift from pondering the undecidable to representing deceptively simple operations in machine actionable form that we can apply billions and billions of times. While we will continue to ponder the meaning of concepts such as “justice” and “goodness,” we now need systems that can reliably distinguish “iron” as metal from the verb by which we press clothing. In classics, we could use a lexicon with more up-to-date information of the various meanings of the Greek word *ἀρχή*, but we need systems that can, with reasonable accuracy, distinguish where Greek *ἀρχή* corresponds more closely to English “beginning” or “empire.” 24

The introduction to this collection has already called for a Cyberinfrastructure, including both collections and services, that can make an ever increasing body of knowledge about the Greco-Roman world intellectually, as well as physically, accessible to an ever widening global audience, supporting many languages and cultural backgrounds. To accomplish this goal, we need not only clever software and well-curated knowledge sources but vast collections from which we can harvest increasingly larger amounts of machine actionable knowledge. 25

Classics and Cyberinfrastructure

The articles in this collection document a range of efforts, each of which is farther along today because of Ross Scaife’s patient and indeed loving support. We see no field within the humanities that has either made the material progress towards — or, even more important, fostered a community to develop and then use — infrastructure on which all of the humanities must depend in a digital world. In this section, we outline a plan forward and argue that any Cyberinfrastructure for the humanities as a whole should begin with classics. 26

The center of gravity for intellectual life in every developed or developing society is now digital and humanity has already begun to arrange an infrastructure around that new center. The term Cyberinfrastructure, however, emerges from the National Science Foundation (NSF) of the United States and it was the NSF that funded the workshop from which this collection emerges.[12] We therefore begin this section by engaging with a discussion that may at first appear peculiar to the United States. In fact, our argument applies as well to Europe, China and every nation, large and small: if we are to prosper in the present, we must better understand the pasts of every community with whom we come in contact. Each nation needs Cyberinfrastructure that can not only preserve, augment, and export ideas about its own cultural heritage but that can import and make as intellectually accessible as possible the cultures, histories, and languages from the rest of humanity. Europe may preside over more languages and a longer historical record, but Europe is not the world. Nor is China, or India, or the Middle East — or all of these together. Every point on the globe is connected. Every cultural community must be prepared to interact with every other, whether history has, for better or worse, already bound them together for millennia or the contingencies of history have kept them apart. 27

Within this larger context Greco-Roman antiquity provides a logical starting point for development. Several reasons stand out: 28

First, Greco-Roman antiquity provides a cultural heritage that is fundamentally international. The Greco-Roman world physically stretched from Ukraine to Spain, from Morocco to Iraq, and from England to the Sahara. Intellectually, the Greco-Roman world provides a foundation for the entire Western Hemisphere. The two largest entities within this space, the United States and the European Union, must collaborate with each other and with every other group that can contribute. A focus upon Greco-Roman antiquity can thus balance the focus upon cultural heritages for which particular nation states must take responsibility. In the United States, we run the risk of replicating in our cultural infrastructure the Anglophone, geographically isolated, culturally leveling tendencies of our history and not preparing for the multi-lingual, physically interconnected, culturally complex world in which we actually live. Any Cyberinfrastructure for classics should draw seamlessly and naturally upon resources scattered across the globe. 29

Second, though this collection has focused primarily upon the textual record, the vast body and variety of data about the ancient world come from 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. Individual nations will be best able to document the physical remains within their borders by integrating locally produced data in international networks of interoperable data. Cyberinfrastructure for Greco-Roman antiquity provides strong, constructive motives for individual ministries of cultures and similar institutions to think globally as well as locally. 30

Third, beyond the influence of any one nation there exists today a finite textual corpus that has exerted and continues to exert, directly and indirectly, an immense influence upon human life. Much of this textual corpus and an increasing body of machine actionable knowledge associated with it is already available under open licenses. 31

Fourth, Greco-Roman antiquity demands a general architecture for many historical languages. Even if we focus upon Greek and Latin, once we begin to contextualize these languages, we will find that we need to work with materials about the ancient near east of which Greece was one component and thus with languages such as Sumerian, Akkadian, Hittite, Old Persian, Coptic and Hebrew. As we consider the reception and influence of Greco-Roman culture, we must work with Syriac and Arabic, as well as with every language of Europe. To work with so many historical languages, we must develop an architecture that can integrate language specific content and services with general services. While we may focus initially on the languages and cultures of the Mediterranean and the Near East, these subjects, daunting as they may be, provide only a component of an environment that must include the historical languages and cultures of the Indian subcontinent, Asia and the rest of the world. 32

Fifth, contemporary classical scholarship is multilingual. Many scientific disciplines manage the language problem by concentrating their publications in English. North American and European classicists alike are conventionally responsible for anything written in, as a minimum, English, French, German and Italian, while classical scholarship appears in Spanish, Modern Greek, Russian, Croatian, Dutch, and any other language spoken by classical scholars. Technologies such as cross language information retrieval (CLIR) are well-established and would be essential in a field such as classics, where scholars want to pose queries in one language to retrieve results in at least four modern languages for which they are officially responsible.[13] Classics is one of the most fundamentally multilingual intellectual communities in the academy and provides the best humanities community within which to explore genuinely multilingual infrastructures. 33

Sixth, our knowledge of the Greco-Roman world casts light upon residents of areas that were at some point part of the Greco-Roman world who are not professional academics. We have natural audiences who speak not only every language of Europe but Arabic, Farsi and Turkish. We must address the challenges not only of professional academics with extensive linguistic training in a handful of languages but of general audiences as well. 34

Seventh, classical scholarship begins the continuous tradition of European literature and continues through the present. Classicists have in recent years led projects on topics such as the history and topography of London, multitexts of Marlowe and Shakespeare, the history of science, 19th century newspapers, and the American Civil War. These have provided us with tangible grounds to argue that the problems of classical studies raise a superset of issues that appear in the humanities before the rise of time-based media such as films and sound. 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. A major classics development project should have allied projects, sharing the same infrastructure in representative domains (e.g., the History of Science, early modern studies, 19th century Anglo-American history and literature). 35

Eighth, classicists have already devoted a generation to developing collections and services. They need a more robust environment and are ready to convert project-based efforts into a shared, permanent infrastructure. They have begun to 36 outgrow the physical systems which they can, as projects, reasonably support. We thus shift discussion to the collections and the services that have already been developed to describe what is now feasible in this field.

Services for eClassics

Services define what we can accomplish. We develop collections in conjunction with services — even if that service consists solely of a mechanical lookup (e.g., call up a particular passage by chapter and verse). We cannot call up Homer, *Iliad*, book 9, lines 44-48 unless we have a digital text and structural markup for books and lines of Homer's *Iliad*. Backend services capture those processes that are available automatically for all textual materials. Every classification service implies both browsing, search and visualization services: i.e., if we identify commentaries among OCR-generated text, we can search for all commentaries; if we recognize that *fecit* is a form of *facio* (Latin "to do, make"), then we can query *facio* and retrieve *fecit*. They provide data on which customization and personalization services draw and to which users respond with corrections and additions. Much fundamental work remains to be done on discovering and perfecting services relevant to classical studies that technology already enables. Ultimately, the decisions become social — technology establishes what is possible but only those engaged in the study of classics can assign, whether by conscious decision or default action, relative values to the services that could be built. Nevertheless, after decades of collection development within the field of classics, a number of services have begun to emerge, some of them actively used for years. The following services represent a core set, and should be components in any cyberinfrastructure for classical studies.[14] 37

The following list offers a minimal set of services, each of which can be built with the technologies available today and each of which addresses established problems relevant to classicists in particular and many humanists. The services below largely address the problem of classification, i.e., applying a set of criteria to find and/or to label materials. Different annotation tasks admit of different levels of certainty: human readers can identify the correct transcription for print on a modern page but lexicographers will disagree on the senses of a given word. Nevertheless, these services aim at more or less deterministic, right-or-wrong answers. We do not include below clustering and other techniques that can detect patterns that require new categories. The services below reflect basic tools on which more open-ended research depends. 38

Canonical Text Services (CTS)

Canonical text services allow us to call up canonical texts by standard chapter/verse citation schemes. Christopher Blackwell and Neel Smith, working in conjunction with Harvard's *Center for Hellenic Studies* (CHS), have developed a general protocol for canonical text services that provides essential functions for any system that serves classicists — or any scholarly community working with canonical texts.[15] Early modern books or MSS that defy current OCR technology can be indexed by conventional citation (e.g., this page of the Venetus A manuscript contains the following lines of the *Iliad*). 39

Optical Character Recognition and Page Layout Analysis

Transcription captures the keystrokes. Page layout analysis captures the logical structures implicit in the page.[16] These logical structures include not only header, footnote, chapter title, encyclopedia/index/lexicon entry etc., but more scholarly forms such as commentary and textual notes. All disciplines have used tables to represent structured data and we need much better tools with which to convert tabular data into semantically analyzed machine actionable data.[17] Much of the work in the Mellon funded Cybereditions Project will focus on this stage of the workflow, focusing on the problem of mining highly accurate data from OCR output of scholarly editions in Greek and Latin. 40

Morphological Analysis

Morphological analysis takes an inflected form (e.g, *fecit*) and identifies its possible morphological analyses (e.g., *3rd sg perfect indicative active*) and dictionary entries (e.g., Latin *facio*, "to do, make"). David Packard developed the first morphological analyzer for classical Greek, *Morph*, over a generation ago.[18] Gregory Crane began the initial work on what would become the core morphological analyzer for Greek and Latin in Perseus in 1984. Neel Smith and Joshua 41

Kosman, then graduate students at Berkeley, extended this work and created a library of subroutines that remain part of the current code base for Morpheus. Morpheus is written in C, has been compiled on a range of Unix systems over the course of more than twenty years, and contains extensive databases of Greek and Latin inflections and stems. Of all the classics specific services with which we are familar, Morpheus is the most mature and well developed. The goal has long been to create an open source version of Morpheus. Desiderata include new documentation, modern XML formats for the stems and endings and a distributed environment whereby users can add new stems and endings.

Syntactic Analysis

Syntactic analysis identifies the syntactic relationships between words in a sentence; it allows us to provide quantitative data about lexicography (e.g., which nouns are the subjects and objects of particular verbs), word usage (e.g., which verbs take dative indirect objects? where do we have indirect discourse using the infinitive vs. a participle vs. a conjunction?), style (e.g., hyperbaton, periodic composition), and linguistics (e.g., changes from SOV to SVO word order). Even relatively coarse syntactic analysis can yield valuable results when applied to a large corpus: working with our morphological analyzer and a tiny Latin Treebank of 30,000 words with which to train a syntactic analyzer, we were able to tag 54% of the untagged words correctly, but the correct analyses provided a strong enough signal for us to detect larger lexical patterns.[19] More robust syntactic analysis based on very large treebanks can yield accuracies of 80 and 85%. Human annotators can build upon preliminary automated analysis to create treebanks, where every word’s function has been examined and accounted for. Treebanks provide not only training data for automated parsing but also explanatory data whereby readers can see the underlying structure of complex sentences — a valuable instrument to support interdisciplinary researchers from fields such as Philosophy or the History of Science who are not specialists in Latin and Greek. 42

Word sense discovery

Word sense discovery automatically identifies distinctive word usage in electronic corpora. Even without syntactic analysis, collocation analysis can reveal words that are closely associated (e.g., phrases such as the English “ham and eggs”) and thus identify idiomatic expressions.[20] Jeff Rydberg Cox developed collocational analysis for the Greek and Latin texts in Perseus and the results are visible as part of the on-line Greek and Latin lexica in Perseus 3.0.[21] Access to translations aligned to the original allows us to identify distinct senses: e.g., *oratio* corresponds both to English “oration” but in other instances to English “prayer.” At Perseus, we have been experimenting with this technique since 2005 and have begun a project, funded by the NEH Research and Development Program, to explore methods for a *Dynamic Lexicon for Greek and Latin*. 43

Named entity Identification

Named entity identification provides semantic classification (e.g., is Salamis a place or a Greek nymph by that name) and then associates names with particular entities in the real world (e.g., if Salamis is a place, is it the Salamis near Athens, Salamis in Cyprus or some other Salamis?).[22] We have developed a serviceable named entity identification system for English and have support from the Advancing Knowledge IMLS/NEH Digital Partnership to extend this work to documents about Greco-Roman antiquity.[23] We expect more general named entity systems to supersede the system that we developed and we are therefore focusing our efforts on creating knowledge sources that will allow these more general systems to perform effective named entity identification on classical materials. Our work focuses on creating (1) a labeled training set, based on print indices, with place and personal names identified, (2) a multilingual list of 60,000 Greek and Latin names in Greek, Latin, English, French, German, Italian, and Spanish, and (3) contextual information, or in other words, which authors mention which people and places in which passages, extracted from the 19th century encyclopedias of biography and geography edited by William Smith. 44

Metrical Analysis

Metrical analysis both discovers and analyzes the underlying metrical forms of digital texts. Metrical analysis provides information about vowel quantity that can improve performance of morphological, syntactic and named entity analysis. 45

Metrical analysis is particularly important for areas such as post-classical Latin, which have very large bodies of poetic materials that will never receive the manual analysis applied to Homer, the Athenian Dramatists, Vergil and other canonical authors.[24]

Translation Support

46 Translation support aims at fluent translation of full text but can provide useful results at a much earlier stage of development. Thus, word sense disambiguation, a component within machine translation, helps translate words and phrases: e.g., given an instance of the Latin word *oratio*, word sense disambiguation identifies when that word most likely corresponds to “oration”, “prayer” or some other English word or phrase.[25] The same service also supports semantic queries such as “list all Latin words that correspond to the English word ‘prayer’ in particular contexts.”

Cross Language Information Retrieval (CLIR)

47 Cross language information retrieval (CLIR) allows users to pose a query in one language (e.g., English) and retrieve results in other languages (e.g., Arabic or Chinese). For classics, CLIR is an extremely important technology because classicists are expected to work with materials not only in Greek and Latin but, at a minimum, in English, French, German and Italian. CLIR is a mature technology where the cross language queries in some competitions perform better than the monolingual baseline systems (e.g., you get better results searching Arabic with an English query than if you searched with Arabic).[26] Classicists should be able to type queries for secondary sources in various languages such as English, French, German or Italian.

Citation Identification

48 Citation identification is a particular case of named entity identification that focuses on recognizing particular: e.g., determining whether the string “Th. 1.33” refers to book 1, chapter 33 of Thucydides, line 33 of the first Idyll of Theocritus or something else? Are numbers floating in the text such as “333” or “1.33” partial citations and, if so, what are the full citations? Primary source citations tend to be shorter and more variable in form from the bibliographic citations found in scientific publications. Perseus has, over the course of more than twenty years, extracted millions of citations from thousands of documents but the citation extractors tend to be ad hoc systems tuned for the subtly different formats by which publications represent these already brief and cryptic abbreviations. In the million book world, we need citation extractors that can recognize the underlying citation conventions of arbitrary documents and then match them to known citations on the fly (e.g, observe numerous references to Thucydides and then infer that strings such as “T. 1,33” describe Thucydides, Book 1, Chapter 33).

Quotation Identification

49 Quotation identification can recognize where one text quotes — either precisely or with small modifications — another even when there is no explicit machine actionable citation information: e.g., it can recognize “arma virumque cano” as a quotation from the first line of the *Aeneid*. The fundamental problem is analogous to plagiarism detection.[27] Support from the Mellon-funded Classics in the Million Book Library study allowed us to begin work on exploring quotation identification techniques.[28]

Translation identification

50 Translation identification builds on both CLIR and quotation identification to identify translations, primary but not exclusively, of Greek and Latin texts that are on-line in large digital collections.[29] These translations may be of entire works or of small excerpts.

Text Alignment

51 Text alignment services most commonly align translations with their source texts and are components of word sense disambiguation systems.[30] Text alignment, however, serves also to create human readable links between source texts

and translations that do not have machine actionable book/chapter/section/verse or other citation markers or between source texts that are tagged with different citation schemes. Text alignment is one of the priorities of the Mellon-funded *Cybereditions Project* at Tufts University.

Version Analysis

52 Version analysis services can collate transcriptions of manuscript sources or of different printed editions of the same work.[31] Version analysis can also be used for automated error correction: when two versions of a text differ and one version contains a word that does not generate a valid Greek and Latin morphological analysis, we flag that word as a possible error and associate the parseable word from the other text with it as a possible correction.[32]

Markup Projection

53 Markup projection services, implicit in many of the services above, automatically associate machine actionable data from one source with the same passage in another source. Thus, an index might state that a reference to Salamis in passage A describes Salamis near Athens but that the reference in passage B is to Salamis of Cyprus. Markup projection services would associate those statements with all references to Salamis in various versions of passages A or B, including not only full scholarly editions but also quotations of those passages that appear in journal articles or monographs

Collections for ePhilology

54 The fifteen basic services described above provide mechanisms whereby human beings can think about the ancient world. Services are dynamic processes that depend upon the algorithmic processing of pre-existing materials. Google and similar comprehensive organizations succeed insofar as they have identified very general algorithms that can generate useful results over thousands of domains to millions of users. Algorithms are the core of computer science. Computer scientists seek to maximize what can be computed and to minimize the pre-existing knowledge that a system needs. In this context, if we can associate 90% of the geographic names in 90% of the English language internet with their locations to which they refer, we may decide that the problem has been solved. Much of the work underway focuses upon such first order approximations which are good enough for many people in many contexts.

55 The remaining 10% or 5% or even 1% may, however, be the space in which the most interesting intellectual work takes place and thus the locus of that value which a digital environment can offer. First, we may be most interested in finding the uncommon instances that are much harder to find. Thus, it is easy to score well on an ambiguous name such as Washington if we are looking for George Washington or Washington state but much harder if we are looking for Washington, MA, or Washington, GA. Second, we need to consider the issues of context. The patterns that we find in English language documents from India and South Africa will, of course, differ from those that we find produced in the US and the UK. If we remain focused on the United States, the 1855 *Harper's Gazetteer of the World* lists more than 150 places named Washington. The early 21st century version of the Getty Thesaurus of Geographic Names (TGN) with which Perseus researchers worked contained only 90 Washingtons. Thus even if we are working with American materials in English but we shift our attention a century and a half into the past, the services optimized for the present rapidly degrade. If we push back into English collections from the 18th century the problem worsens. If we work with early modern documents in English before standardized spelling, the problem grows more complex still. And if we are working with materials in other languages, our generic services may not only degrade but be useless — how may place names can we find in Latin much less Greek or Syriac?

56 Scholarship has always begun where obvious conclusions are not available or, on deeper inspection, prove inadequate. In most cases, readers within a scholarly community can automatically identify the people and places cited by a text but in a small percentage of instances, these references are unclear. Scholars have spent generations trying to decide to which Antonius a particular text refers or which variant reading among the manuscripts (if any) most probably reflects what Aeschylus composed. We may well be able to identify what texts of Plato people have read in dozens of languages over thousands of years and see in a form that we can understand the sorts of things that people have said about Plato as a whole, a particular work of Plato or a particular passage. But such automated analyses and

visualizations provide only the starting point for meaningful interpretation.

In this digital age, a major — and indeed, perhaps the important — portion of our work must center on the space between where the machines can bring us and where our intellectual aspirations lead. As technology advances, some scholarly tasks become wholly automated and are thus obsolete as effective instruments of scholarship. We may print the results of word searches as keywords in context but the production of print concordances is at best a problematic activity: we are better off creating an electronic text and then shuffling the words via various algorithms. If we want to create more sophisticated visualizations, we are better served marking the source text (e.g., identifying each dictionary entry) to create a particular view of that data (e.g., a dictionary organized by dictionary entry rather than inflected form). 57

The following categories of document provide some, though by no means necessarily all, of the foundational data on which we base our work with primary sources. Each constitutes a structured environment through which we human authors communicate with other authors and with automated systems. Each category of document can play the following roles: 58

- **Training data:** Many systems depend upon a training set in which human annotators classify phenomena (e.g., “bank” in passages x, y, z corresponds to a financial institution, but to a river bank in passages a, b, c). Part of each training set is set aside to serve as a gold standard: we test various learning algorithms by training on one part of the training set and then comparing how well it performs on the part that we set aside. Training data thus does not have to be perfect to be useful — in fact, perfection is not a relevant category. In reality, training sets include at least some ambiguous examples and a mature environment must be able to distinguish levels of certainty/community agreement.
- **Corrections and augmentations:** All of the services outlined above include some element of probabilistic analysis. We may be able to identify all variations across multiple versions of a text but still need to refine the ways in which our system classifies differences (e.g., two texts may differ because of an OCR or data entry error rather than because of an editorial change). Our reference works should, insofar as possible, draw upon and then refine and augment an initial automated analysis, thus allowing us to focus time on those instances where we want to change or add to what the machines have done. Our reference works thus provide a place to store our response to the automated work. The audience for these reference works will include both human readers and automated systems which will use the reference works as a training source with which to provide better results.
- **Models and argumentation:** In the end, human authors will continue to analyze, reflect and pose arguments. Print editions, lexicon entries, and even indices of people and places contain models for what an author wrote, what words mean, and what we think we know about the people and places in a document. In a digital environment, these models must be explicit and, where appropriate, encoded into a machine actionable form. Their accompanying arguments must build upon automated methods not available in print when these methods are relevant. We need reasoned arguments and these will retain a familiar expository structure but accompanying data sets may be what have the greatest impact upon intellectual life. The next monographic study of a Greek word, for example, should include annotations that link the findings of and arguments behind that study with the passages to which they are relevant.

The following describe some of the document types that we need in a digital environment. To some extent they all reflect components of comprehensive digital editions and each contributes to the roles that textual data can play in a digital environment.[33] 59

Multitexts

The contribution of Dué and Ebbott in this collection outlines the concept of a multitext. We use the term multitexts here to describe methods to track multiple versions of a text across time. The term multitext does not mean that editors cannot produce their best attempt to reconstruct a source text no longer available to us — we can represent a multitext as a network of versions with a single, reconstructed root. We may well find that the new linguistic and analytical resources at our disposal — especially resources such as treebanks and other categories of linguistic annotation — will 60 allow editors to place old questions on a fundamentally new foundation and to provide new insights into the editions that classical authors produced of their works.

The term multitext does, however, insist upon our ability to track and compare versions over time. In many cases, the original words of an author are as relevant as the Hubble telescope was to Galileo. Petrarch and Machiavelli did not read Teubner Editions or Oxford Classical Texts. We are in a position to begin modeling the texts of our authors as they appeared at different points of time and even the textual universes in which different actors works. Scholars in early modern studies, for example, need systems that can show us at a glance how various sixteenth and seventeenth century editions of classical authors differ from the modern editions that they have laboriously read. 61

First, 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.[34] 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. If the digitized exemplars have associated citation data (e.g., page X in MS Y corresponds to lines M to N of the Iliad, segment A,B,C,D of a given page corresponds to line 38 etc.), then the digital library system can automatically select the page or page segment relevant to a given section of the edition. If that metadata is not present, then the reader will simply have to find the relevant section by flipping through the electronic pages of the witness. 62

Second, multitexts are versioned: they encode not only one reconstructed edition produced by one editor but are designed from the start to represent multiple editions.[35] Any reader should at any time be able to call up visualizations and analyses of multiple editions, seeing which editions are more closely related, which editions had the greatest impact on subsequent editions, which editions are more dependent on particular witnesses, etc. 63

Third, multitexts include multiple apparatus critici, but these apparatus critici are machine actionable. Machine actionable means that textual comments are encoded in such a way that readers can compare the text with readings from MS A vs. MS B and/or select their own readings. While there can be multiple apparatus critici, each apparatus criticus must build upon the same set of common identifiers: a machine must be able to determine that B in one apparatus criticus corresponds to V in another. 64

Parallel Texts

The multitext as described above only covers versions of a text within a single language. In many cases, however, literary texts have exerted their influence in translations that were one or more languages removed from the original. Shakespeare’s worked with Thomas North’s translation of Plutarch, but Thomas North translated Jacques Amyot’s French translation of Plutarch, rather than Plutarch’s Greek. We have to remember that many Greek texts exerted much of their influence when they circulated in Latin or Arabic translation. We need parallel texts of multiple linguistic versions 65

The contribution of Bamman and Crane to this collection introduces the concept of parallel texts and their application to lexicography. Parallel texts can include a single edition and translation (like the Loeb and Budé series) but can also include multiple translations in multiple languages aligned with multiple editions (e.g., an Italian translation of Aeschylus that contains variant translations for a number of major editions). Parallel texts assume some level of common citation schemes: e.g., chapter 86 of book one of Thucydides in an English translation roughly corresponds to the Greek in chapter 86 of book one of Thucydides in standard editions. The more numbered sections, the more precisely citation schemes can align source texts and translations. Parallel text analysis and automatic alignment software can, however, discover many instances where words in the translation correspond to words in the source text. Even if we restrict ourselves to high probability correspondences, we can align our texts far more closely than any traditional citation system. Put another way, once we have page sized chunks of text and translation aligned, automatic alignment can do 66 a better job than manually added structures such as section markers. Such section markers are probably most useful for human readers who want to extract logical chunks. Automatic alignments would be familiar to those who work with Plato and Aristotle, where editions use the page breaks and page sections of particular editions rather than the logical structure of the text itself.

Once we have established the correspondences between different linguistic versions of the text, we need automated methods to help identify likely locations where those versions diverge, whether because a translator misunderstood the original or because the idea of translation was looser than that of later periods. Finally, we need methods whereby scholars can annotate these differences according to the patterns which they determine are significant. 67

WordNets and Machine-Actionable Dictionaries

The contribution of Bamman and Crane in this collection also introduced some of the possibilities for dynamic lexicography in a digital environment. WordNet and EuroWordNet are pragmatic examples of semantic networks, associating words with similar meanings into hierarchical classes.[36] WordNet in particular has emerged as a major tool within computational linguistics and similar resources for Greek, Latin and other historical languages would be an important contribution.[37] Machine actionable dictionaries may resemble traditional lexica in format but differ in that they contain far more citations than could ever be printed, they can be updated continuously, and their information is from the start structured to support morphological, syntactic and semantic queries. True machine actionable dictionaries must articulate word senses in such a way as to help both human and machine readers to recognize these senses as precisely as possible. 68

Treebanks, Linguistic Annotations, and Machine-Actionable Grammars

Treebanks are databases that label the syntactic role of each word in a set of sentences. These syntactic tags constitute parse trees (hence the name) that can be used to analyze lexical, syntactic and even rhetorical patterns.[38] Treebanks tend to have fairly compact tagsets — they might not encode purpose clauses per se but allow users to query for patterns such as *ut* followed by a subjunctive. 69

Syntax is important but by no means the only subject of linguistic annotation. Co-reference annotation maps pronouns to their referents (e.g., “he” in passage X refers to Julius Caesar). Annotation languages have emerged to capture higher level semantic phenomena such as temporal expressions (TimeML).[39] 70

We use machine actionable grammars to describe resources comparable to print grammars. These may have hundreds or thousands of observations, each roughly corresponding to the numbered paragraphs of their print predecessors. But in a machine-actionable grammar, each paragraph would include not only citations but a set of patterns (e.g., *ut* heading a subordinate clause followed by the subjunctive) and some indication of the precision (how many false hits the pattern would retrieve) and recall (how many correct hits the pattern would miss). The machine-actionable grammar would thus build on the treebank. Where the treebank would stress use of a smaller number of categories to describe the relations of individual words, machine readable grammars would suggest an open-ended set of more complex phenomena inferred from the corpus. 71

Machine-actionable indices of people, places, organizations, etc.

The contribution by Elliott and Gillies in this collection outlines the major issues surrounding geographic information in classical studies. We also need to represent information about people, organizations, technical/scientific terms and other entities with regular features. 72

The underlying principal of machine actionable indices is the same as that of their print antecedents. Machine actionable indices differ in at least two ways. First, the structure of the index entries is explicit: we can extract headwords, hierarchical structures (e.g., “Athens, (1) Religion …. (2) Government …”) descriptive labels (e.g., “born at X,” “stood for consul in Y”), and associated citations. Second, index headwords contain the most general possible identifiers. Thus, we don’t simply cite Athens, Greece, or Thucydides the Historian, but add the identifiers such as the 73 numbers for Athens (TGN 7001393) and Thucydides (TLG 0003) in the *Getty Thesaurus of Geographic Names* (TGN) and the *Thesaurus Linguae Graecae Canon* (TLG) respectively.[40]

Propositional Knowledge

Propositional knowledge includes standard database fields: e.g., author=Thucydides + Title=History-of-the-Peloponnesian-War in effect states that Thucydides is the author of the *History of the Peloponnesian War*. Propositional data is, however, designed to support reasoning: e.g., if two people share the same two parents, then we can infer that they are also siblings; if someone was born after an author died, then the works of that author cannot refer to that person. 74

Such propositional reasoning rapidly becomes computationally complex. More significantly, the underlying propositions rapidly become idiosyncratic, as each observer creates slightly different categories and our propositional knowledge becomes internally inconsistent — as soon as computer scientists began converting print reference works such as the *Oxford English Dictionary* to digital form, they discovered that human editors were never fully consistent.[41] 75

The Historical Event Markup and Linking (HEML) which Bruce Robertson describes in his contribution to this collection illustrates the measured use of an ontology to do a great deal but not too much — HEML did much to shape the newest extensions in the Text Encoding Initiative (TEI) methods for representing named, dates, people and places.[42] If we restrict ourselves as much as possible, however, to established ontologies (a common set of propositions), then we can build off the work of others. Insofar as we can share the same ontologies with broader communities, we have a chance to create propositional knowledge that can be integrated with propositional knowledge from other sources, creating a much larger and more powerful knowledge base than any single project could develop. Put another way, a large number of propositions describing a finite set of well-defined phenomena will probably yield far more useful results. Individuals may extend shared vocabulary with their own categories but retain a common set of categories by which at least part of their data can interact with other systems. All of the reference works listed above depend upon propositional knowledge of the form “A has property B”: the string *“Arma virumque cano, Troiae qui primus ab oris”* has-citation Vergil-Aeneid-book-1-line-1; *fecit* has-language Latin and *fecit* has-morphological-analysis; archê-in-passage-X has-sense “empire.” A treebank contains compound propositional statements such as *agricola* is-a noun and *agricola* is-subject-of *fecit*. We include propositional knowledge as a separate category to emphasize categories not included above. Thus, the CIDOC-CRM ontology includes a wide range of categories for art and archaeological objects and HEML provides a vocabulary for describing people, places and events in time.[43] 76

Commentaries

A true digital commentary must build judiciously upon all of the tools listed above. Full commentaries should include annotations identifying every phenomenon of interest to its intended audience: every word should be morphologically disambiguated, every sentence should have its syntactic data encoded; every major variant should be labeled; every person and place should have at least one identifier from a general work or a label indicating that this is a place/person/institution not yet in available reference works and a new identifier. Put another way, if scholars have developed a widely recognized classification scheme (word senses in a lexicon, numbered paragraphs in a standard grammar, metrical analyses), then fully commented texts will have categorized every instance of each relevant phenomenon in a text. And, of course, commentaries must from the start allow commentators to include variant explanations for the same phenomenon (e.g., proposographic disputes about which Antonius is meant, textual arguments about which reading is correct). 77

Publication for a Cyberinfrastructure

> An Athenian citizen does not neglect the state because he takes care of his own household; and even those of us who are engaged in business have a very fair idea of politics. We alone regard a man who takes no interest in public affairs, not as a harmless, but as a useless character; and if few of us are originators, we are all sound judges of a policy. The great impediment to action is, in our opinion, not

> discussion, but the want of that knowledge which is gained by discussion preparatory to action. (Thuc. 2.40.2, after Crawley)

For us, public affairs go beyond the individual decisions of a particular government but extend to all discussion. We may be professional academics, privileged to earn a living by working on the subjects to which we have dedicated our lives, but we enjoy that privilege because we serve the broader interests of humanity. Our work within the academy is only a means towards the greater goal of supporting intellectual life and the general understanding of the past. 78

Before discussing some of the essential features that characterize true publication in a digital age, we distinguish, in the context of this discussion, archives and libraries. For our purposes, libraries provide the foundation on which public discourse takes place. Libraries constitute the most advanced and efficient space with which society is able to conduct discourse that extends across time and space and that depends upon preservation of, and access to, the terms of discussion. 79

Archives, Libraries and Intellectual Discourse

> He had also, says he, such a library of ancient Greek books, as to exceed in that respect all those who are remarkable for such collections; such as Polycrates of Samos, and Pisistratus who was tyrant of Athens, and Euclides who was himself also an Athenian, and Nicorrates the Samian, and even the kings of Pergamos, and Euripides the poet, and Aristotle the philosopher, and Nelius his librarian; from whom they say that our countryman Ptolemaeus, surnamed Philadelphus, bought them all, and transported them with all those which he had collected at Athens and at Rhodes to his own beautiful Alexandria. (Athenaeus, *Deipnosophistae* 1.1, tr. Yonge)

Our varied conceptions of a library are both descriptive and prescriptive: these conceptions shift as material culture changes the methods with which we can manage information. In the Greco-Roman world, Alexandria had the most famous library and every lover of Greek literature sighs to think of the tragedies of Aeschylus, Sophocles, and Euripides, the poems of Sappho and the other works that once lay among its holdings and are now lost. The library at Alexandria was based upon miraculous technologies such as papyrus production and sea-born travel as well as writing.[44] 80

Popular conceptions of institutions such as libraries evolve along with the capabilities of their enabling technologies. The ancient library at Alexandria was not the instantiation of a Platonic ideal but the best use of the most advanced methods of the time. The library at Alexandria brought texts from around the Greek world into a single location. In the industrialized world, we have used industrialized print technologies to create hundreds of large libraries around the world, in effect protecting long-term access by maintaining multiple copies of the same work in widely separate locations. In the digital world we can not only create far more numerous copies and greater redundancy but our libraries are no longer inherently limited to physical locations.[45] They can at any point reach any point on the earth. Twenty-first century collections become libraries only insofar as they fulfill the need to provide access over time and across space. Long term preservation and global access are foundational challenges for our new information infrastructure.[46] 81

The passage quoted attributes to an intellectual of the second century CE the claim that he had assembled an unparalleled collection of ancient Greek books. Two features from the underlying Greek are worth noting. First, no word corresponding to “library” actually appears: the Greek phrase (*bibliôn ktêsis*) describes the “possession of books” and does not designate either a place or an organization. Second, the passage above speaks in terms of individuals and collectors. The one exception, Nelius, is not a librarian: the Greek text probably includes an error but the term applied to Nelius (*diatêrêsanta*) states that he preserved the books of Aristotle and does not designate a generalized occupation such as the term librarian implies. We have left the nineteenth century translation unchanged to illustrate how easily we all project the categories of the present into sources from the past. 82

A collection of hand-written documents, however, did not fit the dominant conceptions of libraries that took shape in print culture. We still call the ancient manuscript collections of Europe libraries because they bore this name, but in the massive libraries that emerged in the 19th century manuscripts, pamphlets and everything that did not fit the exacting 83 demands of academic publication was preserved in special collections and archives. There, these documents would await the scholar who would cull them for information or create printed editions of them that could circulate and play an active role in the mainstream of intellectual life. For each surviving ancient text of Greek and Latin the *editio princeps*, the first printed edition, no matter how problematic its contents, represented a milestone and a new birth, marking the transition from handwritten manuscript into the new technology of print. Works still available only in manuscript were, in print culture, the material for published editions and printed facsimiles. They had not yet been published in print and thus were not yet a part of the citable record upon which general human discourse could depend.

In the past decade, the academic library system has quietly shifted again. The print libraries of the 19th and 20th century have, in effect, become the archives of the 21st century, as publication and discourse in the most heavily supported disciplines have shifted entirely to a digital medium. The debate about print and digital information may continue but the infrastructure of mainstream intellectual discourse is now digital. The hotter the scientific discipline, the shorter the half-life of its publications — the last five or ten years of published material is enough to support many and probably most cutting edge research projects. Biologists studying changes in flora and fauna need access to as much historical data as possible — for them observations from the 18th century provide foundational data. The *Biodiversity Heritage Library* may be the last major historical collection to be digitized within the sciences.[47] With this project, the last major community of scientists is leaving the print world — and even these scientists maintained their own separate print library infrastructure: all ten of the institutions participating in the Biodiversity Heritage Library draw on specialized libraries that were already distinct from the libraries upon which humanists depend (e.g., the Harvard University Botany Libraries rather than Widener Library).[48] The disciplines in which the advanced nations invest the most now, in effect, print what they need on demand. 84

But just because information is on-line does not mean that that information has exploited the full potential of the digital medium. The debate has shifted instead to the question of open vs. closed access. The extraordinary cost increases for scientific journals have done more than anything else to drive the principle of open access — roughly one quarter of the entire acquisition budget for the Tufts University library in 2007, for example, went to a single scientific publisher, which does not invest any significant sums in the research that it publishes.[49] In 2008, decades of rhetoric finally led to action.[50] Even under a pro-business Republican administration, the status quo has been intolerable. In April 2008, the National Institutes of Health (NIH) instituted an open access legal mandate that all publications produced with NIH support be deposited in the open access PubMed repository within twelve months of publication.[51] 85

The massive library collections at Harvard University have been a magnet for scholars and the university has traditionally been quite conscious of the investment it has made and the advantages which that investment confers upon it — the Boston Library Consortium is often described as “everyone but Harvard.” Nevertheless, Harvard University surprised many observers by taking a dramatic stance in favor of open access. The Faculty of Arts and Sciences at Harvard University voted in February 2008 “to give the University a worldwide license to make each faculty member's scholarly articles available and to exercise the copyright in the articles, provided that the articles are not sold for a profit.” [52] The ruling automatically applies to all faculty publications and individuals must “request a waiver of the license for particular articles where this is preferable” — faculty cannot, according to the language of the press release, simply refuse to exempt themselves but must request waivers on a case by case basis. Steven E. Hyman, Provost at Harvard University framed the new policy in terms of responsibility: “The goal of university research is the creation, dissemination, and preservation of knowledge. At Harvard, where so much of our research is of global significance, we have an essential responsibility to distribute the fruits of our scholarship as widely as possible.” Harvard is, of course, only a single institution but the actions of its faculty and administration provide a powerful example of how conventional thought has begun to shift. 86

Google may ultimately solve the problem of access to the earlier print record. Through its Google Books project, Google has already digitized millions of books (and a striking amount of 19th century classical scholarship).[53] The University of Michigan, for example, has entered into a partnership to digitize the entire print collection of the University Library — collections which “number over 7 million volumes, covering thousands of years of civilization, from papyri to reports of 87 the latest advances in science and medicine.” [54] The legal agreement between Google and the University of Michigan contains a clause entitled “searching free to the public” that asserts that Michigan content be made available at “no direct cost to end users.” [55] Google is not asserting open source — Google does not allow commercial competitors to build services on top of the books that it paid to digitize and legal issues remain to be resolved. Nonetheless the logic behind the vast Google digitization effort moves academia much farther towards open access for a global audience.[56]

88 Classicists have already begun taking steps to make their core primary materials available in the interoperable formats and open licenses needed for teaching and research in a digital world. The Perseus Digital Library released the TEI-compliant XML source files for all of its primary sources and accompanying translations in March 2006 under a Creative Commons license. Harvard’s Center for Hellenic Studies (CHS) has also undertaken to extend this effort and announced in August 2008 a plan to create a digital library of new TEI-compliant XML editions for the first thousand years of classical Greek, including “at least one version of every Greek text known to us from manuscript transmission from the beginning of alphabetic writing in Greece through roughly the third century CE.” [57] Support from the Mellon Foundation has allowed Perseus to begin building a comprehensive collection of scanned critical editions for every major Greek and Latin author. The initial results of this work are already available for public download at the Internet Archive and under a license that allows anyone to create new derivative works using their own OCR or text mining software and publishing the results in their own services.[58]

89 If we are to understand what form we would like our libraries to assume, we must first consider what we expect from the publications that will populate these libraries.

Features of Publication in a Digital World

Socrates:	And every word, when once it is written, is bandied about, alike among those who understand and those who have no interest in it, and it does not know with whom to speak or not to speak; when ill-treated or unjustly reviled it always needs its father to help it; for it has no power to protect or help itself.

Phaedrus:	You are quite right about that, too.

90 Scholars have written about the ancient world since antiquity itself, and we build upon more than half a millennium of the scholarship that print made possible. A great deal of material about the Greco-Roman world exists in digital form, but only a small subset of that material can fulfill its potential in a digital world. The essential criteria for true publication are different in the digital world because the digital world supports services that are not feasible in print and can reach audiences millions of times larger than academic print publications could reach. The fact that a resource exists in a digital format is a necessary but not sufficient condition: just because an object of potential relevance to classics is digital does not mean that it is useful.

91 Not only the print volumes that sit upon our library shelves but the digitized publications to which commercial entities sell access have all become, within the digital world, archival materials, tied to a few discrete points on the earth and membership in specialized organizations. Whatever the merits of their content, these essays are important because, despite the vast body of existing scholarship, these essays are among the first original works of classical scholarship to meet the minimal criteria for publication in a digital age.

92 Scholarly publication in a digital age must satisfy at least the following four conditions. These four conditions overlap, of course, with those familiar from five centuries of print culture, but, of course, they also must adapt to the digital foundation on which all shared intellectual expression already depends.

93 First, the content must be of interest to someone other than its producers. In academia, we have developed peer review as an instrument to assert that a particular intellectual production has sufficient value to warrant a permanent place in the scholarly record and we used traditional peer review in this collection as well. Peer review is, of course, no guarantee — and readers will come to their own conclusions about what is published here, as they do about everything

that they read. Other models exist to achieve the same goal and we should not confuse the instrument of peer review with its purpose.[59]

Second, the content must be in a format that we can preserve and use for long periods of time. Print culture developed for the organization of books and articles conventions that have proven so successful as to become almost invisible: we take tables of contents, chapters, footnotes, indices, bibliographies and other conventions for granted. In a digital environment, machines are the first and essential readers of all published materials — where more is written than any one person can digest, we depend upon what machines can extract to identify those few objects on which we can focus the limited attention and intellectual capacity of the human brain. The articles in this collection express their basic structures in a standardized format that machines can understand. More sophisticated documents will surely emerge but these are likely to enhance, rather than abandon, the structures within this collection. By investing in the XML markup we have conformed to the best practices of the present so that the digital librarians in future generations can manage these articles within their digital collections. 94

Third, the content must have at least one reliable long-term home. In print publication, authors needed publishers to put their work into circulation. Publishers committed, however, only to provide very short-term access. Preservation in print culture has always been the task of libraries. Even if war or natural catastrophe destroyed one library, other libraries preserved separate copies of each work and these could be reprinted or reproduced with increasing facility. In a digital age, distribution is trivial — any web page could in early 2008 reach more than half a billion machines.[60] Preservation is, however, a major challenge. Classicists know that they can usually track down copies of the most obscure 19th century dissertations somewhere because libraries have worked hard to preserve academic publications. A 2005 study found that half of the URLs cited in a 1995 issue of *D-Lib Magazine*, a major venue for publication, no longer worked [McCown 2005]. Libraries have, however, moved to address this situation and have created institutional repositories with which to fulfill in this new digital world their ancient mandate of preserving what they collect.[61] The articles in this collection are part of the permanent collection of the University of Illinois at Urbana Champaign (UIUC) libraries. 95

Fourth, the content can circulate freely — it is, indeed, truly public and thus published. A decade ago, this idea was radical and unnerving to many of us, but the Stoa Publishing Consortium always supported open access from its creation in 1997. In the quotation that opens this section, Plato’s Socrates expresses anxiety that information, once represented in a physical medium is separate from its producers and begins a life of its own. In the end, we have overwhelming reasons to leave these anxieties behind. First, we need both our primary and secondary sources to be open for analysis by as many systems as possible if we are to exploit the full power of the digital world and to fulfill our professional obligations as scholars. Second, each scholar, department, discipline, college, and university is, at some level, locked in a Hobbesian war of all against all. College and university web sites are very expensive to produce and maintain but they are freely accessible because each institution is competing for exposure. Subscription revenues do not pay for scholarship. Third, we have plenty of money in the system to pay the costs. During 2005, the 123 members of the Association of Research Libraries invested more than 1.1 billion dollars in their collections.[62] Our interest lies in maximizing exposure. We need to shift from importing the products of third parties and towards exporting the productions of our scholars, departments, and disciplines. Some of the authors in this collection remember hearing that it would be impossible for libraries to provide access to electronic materials — they didn’t have enough resources to collect print. Likewise, we heard that universities could never support web sites — they were too expensive and the budget was already overstressed. Nothing can ever change — but everything always does in the end. The first three reflect a narrowly construed Hobbesian model of self-interest, but they all support the fourth and most important reason. We have a moral obligation as scholars to preserve, expand and disseminate, as broadly as possible, as much of the human record to as much of humanity as possible. For this reason, we have adopted a Creative Commons license not only for the publications in this collection but for all of our work. 96

Peer review, the *Digital Humanities Quarterly* (DHQ) XML style-sheet, institutional repositories and Creative Commons licenses are the four instruments by which we address ideals of content, form, stability and openness inherent in true digital publication. 97

The Scaife Digital Library (SDL)

> The advice of Themistocles had prevailed on a previous occasion. The revenues from the mines at Laurium had brought great wealth into the Athenians' treasury, and when each man was to receive ten drachmae for his share, Themistocles persuaded the Athenians to make no such division but to use the money to build two hundred ships for the war, that is, for the war with Aegina. This was in fact the war the outbreak of which saved Hellas by compelling the Athenians to become seamen. The ships were not used for the purpose for which they were built, but later came to serve Hellas in her need. (Herodotus 7.144, tr. Godley)

Themistocles somehow convinced his fellow citizens to forego a windfall payment and to invest instead in a navy. Even then, the nominal object of the navy — a war with the nearby island of Aegina — masked the vastly greater, but inconceivably distant, Persian threat. Aegina looms as a presence visible from the Acropolis. Herodotus elsewhere (Hdt. 5.53) reports that the Persian capital at Susa was a three-month journey from Ephesus on the West coast of modern Turkey. 98

While most of us remained focused upon publishing our own work under our own name and building digital resources that would serve our own projects, Ross Scaife early realized that there were bigger issues at stake than a few drachmas of scarce prestige in a small academic field. The idea behind the Scaife Digital Library (SDL) reflected Ross's own long-term interests: a 1997 grant from the Fund for the Improvement for Postsecondary Education helped Ross Scaife found the Stoa Publishing Consortium to pioneer new models of publication to enhance learning and intellectual life.[63] 99

The SDL is a new, virtual collection designed to support the digital publications that meet the four criteria outlined above. The first plans for the SDL were presented at the beginning of a two day workshop on "What do you do with a million books?," Humboldt University in Berlin on March 17, 2008, two days after Ross Scaife died in Kentucky. On August 6, 2008, the Institute for the Study of the Ancient World, based at New York University, funded a planning meeting hosted at Harvard's CHS in Washington, DC. The first release of the SDL was announced on November 6 of the same year, at the TEI Annual Meeting at King's College London. 100

The SDL contains durable digital objects that satisfy the four criteria of digital publication outlined above: 101

1. The content has been judged worth preserving. Peer review is the most established mechanism to establish this judgment.
2. The content is in a defined, approved format suited for preservation over long periods of time. Examples include XML documents encoded according to the Guidelines of the TEI and of EpiDoc.[64]
3. Each object has a long-term institutional home separate from the individual or group that produced it. Digital repositories at Brown University, NYU, Tufts University, and UIUC among other institutions currently store the initial objects in the SDL.
4. Each object is available under an open license. Where authors create documents to present a particular scholarly voice at a particular time, an open *access* license should allow third parties to quote and republish the document but not to change its content. Where authors create works designed to encapsulate general and evolving points of view (e.g., lexica, commentaries, editions), then an open *source* license is necessary so that third parties can, in fact, modify the content. In this case, versioning systems track and identify who was responsible for each change.

The SDL is simultaneously an idea, a concrete collection, and an organization to produce new content. Any digital objects that satisfy the four criteria of publication automatically belong to the SDL — thus every article already published by the DHQ can be treated as part of the SDL because each DHQ article satisfies all four criteria. Ross Scaife was a classicist and classics offers the initial center of gravity for the SDL, but we exclude nothing relevant to the humanities.

The SDL is also a concrete collection: it includes a catalogue of known objects and the information needed for automated services to collect each digital object from its home repository. We hope to see objects from the SDL in a 102 range of locations and organizations: with Internet giants such as Google, at particular computational and storage Grids, and on local computing clusters.

103 Finally, the SDL is an organization designed to produce new content. The production of new SDL content can be a simple decision that any digital object produced by a particular third party (e.g., *DHQ*) automatically becomes part of the SDL — in this, the SDL mirrors the standing subscriptions by which libraries traditionally purchased every publication from particular publishers in print culture. But the SDL, however, also provides editorial review of original content.

104 The SDL does not, however, provide services for end users. The SDL may include the code for those services that only humanists can be expected to provide (e.g., an advanced morphological analyzer for classical Greek) but the SDL does not plan to provide those services. The SDL provides a long term home for the objects which others can analyze or make accessible in various systems. We require that each object have an approved format so that as many groups as possible will develop the largest possible number of services with which to make SDL objects useful to the widest possible audience. In addition, we require that each object have a long term home, which in effect, states that we have entrusted libraries to apply their traditional functions of preservation and access for SDL objects. The requirement that each object have an open license reduces our dependence on any one institution: we hope that there will be many copies of each object from the SDL, both under formal preservation systems (such as LOCKSS) and in thousands of informal collections.[65]

105 The SDL thus answers questions of production and preservation but questions remain. The digital environment allows us to rethink not only publication but who can publish and how we divide labor in the scholarly world.

The Work of Scholarship: New Divisions of Labor in the world of Google and Wikipedia

Theban Herald:	Who is the despot of this land? To whom must I announce the message of Creon who rules over the land of Cadmus, since Eteocles was slain by the hand of his brother Polyneices, at the sevenfold gates of Thebes.

Theseus:	You have made a false beginning to your speech, stranger, in seeking a despot here. For this city is not ruled by one man, but is free. The people rule in succession year by year, allowing no preference to wealth, but the poor man shares equally with the rich.

> Master Tyndale happened to be in the company of a certain divine, recounted for a learned man, and, in communing and disputing with him, he drove him to that issue, that the said great doctor burst out into these blasphemous words, “We were better to be without God's laws than the pope's.” Master Tyndale, hearing this, full of godly zeal, and not bearing that blasphemous saying, replied, “I defy the pope, and all his laws,” and added, “If God spared him life, ere many years he would cause a boy that driveth the plough to know more of the Scripture than he did.” [Foxe 1965]

106 The papers in this collection have focused upon the practices of scholarship. In this section we consider the work of scholarship and the associated division of labor. The center of gravity for intellectual life has not only shifted, decisively and forever, to a digital medium but the relative position of professional humanists has changed as well. To some extent, that division of labor has already begun to shift. The scholarly practices to which we award Phds, tenure and promotion may have remained largely unchanged but new practices of intellectual life have exploded onto the scene. Most of us like to think of ourselves as a progressive force, but we, in the eyes of many, more closely resemble the bullying Theban Herald of Euripides’ *Suppliants*. Worse, we may appear to have become like the Athens of Thucydides, a *turannos polis*, a city-state in which only holders of Phds or even those with professional academic appointments alone have the right to speak and contribute. The Tyndales of the twenty-first century maintain blogs, work for the Open Content Alliance (OCA), write for Wikipedia, produce content under Creative Commons open licenses and drive explosive growth of other, novel forms of intellectual production. All of those who have written for this collection feel a profound obligation to

address this gulf between the work that we do as professional scholars and the messy, passionate, unruly, intense streams of activity that have carried *Wikipedia* and other efforts so far.

Professional academics have played, insofar as we can tell, almost no direct role within this historic movement. The authors of this conclusion do not know of any academic who has included Wikipedia along with their conventional publications in their yearly reviews. We do know that, as of the end of August 2008, Wikipedia contains more than two and one half million entries. And we know that this resource has proven astonishingly useful, its flaws real but, when systematically analyzed, no worse than those of conventional, centralized reference works.[66] 107

No one knows how much labor the various language versions of Wikipedia have absorbed — in part because volunteers have contributed the vast majority of the labor and volunteers do not track billable hours. Wikipedia does cost money — the 2005 budget for Wikipedia was $739,200, while the overall Wikimedia foundation reported a budget of 4.6 million dollars for 2007-2008.[67] The aggregate cost will thus repesent well under 40 million dollars (i.e., which would be the cost if they had spent $5,000,000/year each year since 2001 when they began). Clay Shirky, however, recently estimated that Wikipedia represented 100,000,000 hours of labor — thus representing at least 1 billion dollars in labor. The ratio of paid to volunteer labor is thus at least 20 to 1, and probably very much higher. The National Endowment for the Humanities (NEH), by contrast, requested a budget of less than 145 million dollars for fiscal year 2009 — it would take almost seven years of the entire NEH budget to produce $1,000,000,000. The labor power unleashed by this one new mode of intellectual production is extraordinary. 108

Scholarly publications incorporate a great deal of accumulated labor. In classics, the language barriers make such embedded labor relatively easy to identify — classicists need expertise in the Greek and Latin languages, familiarity with the ancient core texts of at least one of these languages, and enough knowledge to work comfortably with book-length studies in English, French, German and Italian. If we consider four years of undergraduate education and six years of doctoral studies as one model of scholarly apprenticeship, each scholarly publication represents years of embedded labor. When a faculty member devotes a month or two in the summer to a new publication, we thus need to consider not only the hundreds of hours invested during that summer but all the years of work on which that scholar is drawing. 109

Wikipedia and other forms of community-driven intellectual production ultimately increase the audience for — and thus the realizable value of — advanced scholarship. Professional academics need to decide how they wish to respond to this vast audience. Many of us are products of a print culture in which our publications simply could not reach beyond a few hundred or, at best, thousand research libraries. We had no reason to write for audiences that our publications would never reach. Furthermore, the professionalized incentives of academia rewarded us for producing work that would impress our colleagues and facilitate tenure, promotion, and other signs of academic success. We now have, however, radically new technologies and social practices with which to advance the intellectual life of humanity as a whole. 110

Twentieth century print culture produced scholarship that required a great deal of training to produce and almost as much training to understand, much less appreciate. We now see a world emerging with much lower barriers for entry. 111

- **Tangible contributions**. Automated methods can do an immense amount but they benefit as well from very large amounts of skilled human labor. Many basic tasks reflect the strengths of human intelligence and provide opportunities for students and non-professionals to contribute tangibly to the infrastructure on which the study of classics depends. The essays by Blackwell and Martin and by Elliott and Gillies document areas in which students can quickly begin contributing tangibly to our understanding of the ancient world. Bamman and Crane describe the emerging role of syntactic databases — treebanks — for the study of classical Greek and Latin. Even if we have a treebank with millions of words already analyzed to serve as a training set for an automated syntactic analyzer, the best automated systems do not, at present, provide more than 87 or 88% accuracy — enough for many analytical purposes but not perfect. Greek classes at Brandeis, Tufts, Furman and elsewhere have already begun to integrate the production of syntactic data into their curricula. The method is straightforward. Treebanks use their tags and methodologies but, in essence, the production of treebanks depends upon ancient practices of reading — we need to identify the main verb, its subjects, objects, etc. Two students can, for example, analyze each sentence, the class can then discuss the points at which they differ, and produce carefully analyzed sentences that may include variant interpretations.

When given a particular set of tags and relationships most readers will agree on the syntactic relationships between most words in most texts, however, some Greek sentences support multiple interpretations, whether because we are not sure what the author originally wrote or because the text that we have reconstructed is fundamentally ambiguous. Ultimately, the syntactic analysis for some words in our surviving texts remains an object of research.

Other tasks that are in most cases straightforward can be the object of research as well: in some cases we cannot determine to which Antonius or Cleopatra a particular passage alludes and we depend upon skilled prosopographical analysis to rank the possibilities. We find place names where we do not know for sure the original location. Word sense disambiguation depends upon the senses that we ascribe to a word and thus upon semantic analysis that can become complex for common words.

We thus see a gradient of tasks. In many cases, students and undergraduate classes can improve upon the results of automated processes and/or provide the initial training data from which, in turn, automated methods can analyze much larger bodies of material. In some cases, the answers to conceptually simple questions (e.g., who is the Antonius in this passage? What is the structure of this sentence?) are not immediately clear and have historically provided scope for some of the most skillful classical scholarship. The patterns visible from the many passages that are not controversial will, when aggregated and analyzed, allow us to place discussions of ambiguous instances on a more explicit and quantified footing. We may even find scholarly consensus advancing as new scholarly instruments, developed in large measure by students and the general public, allow us to shed light upon old problems. Thus, we have a space that provides ample room for contributors at a various levels of expertise.

- **Undergraduate research**. Once we have large databases of information we can begin to see patterns that were not visible before. We rely upon automated methods of analysis to direct our attention to interesting patterns and thus to serve as the starting point for, rather than a conclusion to, analysis. It is important to emphasize that we do not need perfect data to identify major patterns —a recent study conducted by David Bamman showed that even when automated syntactic analysis generated results that were as low as 50% accurate, some significant linguistic patterns were visible despite the noise of a 50% error rate.[68]

  New sources of data open up possible research topics to which our advanced undergraduates can realistically aspire. The Homer Multitext Project, for example, has published high resolution images of the most important manuscript of Homer, the 10th century Venetus A, making visually accessible scholia and readings that have never been published, much less translated. Students are well able to produce initial diplomatic editions with basic contextual information and English translation. Published in standard formats under open licenses and in long term institutional repositories, such works can provide the foundation for a new generation of editions. Generations of students can productively provide the intellectual apparatus needed to understand the detailed page images already being produced in Europe and North America for manuscripts of Homer and other classical authors, fundamentally changing the role that these source materials can play in intellectual life. Likewise, the creation of treebanks allows us to see patterns of word usage, linguistic practice and individual style. Even now, as we develop large automated treebanks, students can create treebanks for individual works and control samples to produce original research: thus, given a treebank and the ability to find Greek words corresponding to English, students could undertake valuable systematic studies that were not practical before (e.g., the semantics of words for “power” in Herodotus and Thucydides). The results of their research can be published through our university repositories, connected to every passage on which they shed light, and preserved, as permanent contributions, long after their youthful authors have passed from the scene.

It would be hard to overstate the possible opportunities of practical undergraduate research for classics and the humanities in general.

The field of classics — and, indeed, every field within the humanities — needs to adapt itself to the challenges and opportunities, some realized, others emergent though visible in outline, that this digital environment has thrust upon us. 118

First, all classicists are digital classicists. Insofar as the practices of our work advance research projects imagined within the limitations and for the tiny academic audiences of print culture, we are antiquarians. We may not believe in particular ideas such as the "judgment of history," but we do believe in conventional ideas and are confident that the implicit assumptions about what constituted scholarship in the twentieth century will give way to new conventional ideas. Each of us working now for an audience in the future is making bets about what those conventional ideas will assume. The authors of this conclusion are not so sanguine as to believe that the culture and languages of ancient Greece and Rome will inevitably flow outwards into the hearts and minds of humanity as vigorously as we hope. Technology constrains and enables the space within which we move. How well and how quickly we in classics and the humanities adapt to the niches within this space depends upon the decisions that we make (however unpredictable the outcomes of those decisions may be). 119

We do not know yet what common technological knowledge classicists must share. We cannot all be accredited system administrators or application programmers. On the other hand, it is hard to accept complaints that the TEI Guidelines or the underlying structures of treebanks are too complicated for scholars who work with six languages. The services outlined above can use textual and syntactic markup to enable new forms of scholarship and of reading support but such data structures are, fundamentally, surface expressions of traditional ideas. Habits from the past and anxiety about the future are the major barriers. Those who have succeeded in the traditional tasks of classical philology will, if they can muster the necessary labor, find themselves in a world that allows them to pursue their traditional tasks more fully. If they can read Pericles' Funeral Oration in the original Greek, they are well able to master any general technological system. 120

Classicists need only to exploit the analytical tools and conventions of intellectual discourse available to them to achieve their goals. For us, the blogs, wikis, assorted web pages and other digital tools simply challenge us to adapt the complementary goals of rhetorical power and intellectual discipline. We hope that others will more fully realize these goals than has been possible so far for us. 121

Second, classicists need some scholars who have more advanced knowledge of the technology. We do not have the resources to sustain a subfield such as bioinformatics, but the broadening textual collections and treebanks now starting to emerge for Greek and Latin build upon many of the same techniques used to find patterns in the human Genome. The most important philologists now at work may well be the classicists who have joined the field of computer science and are now laying the foundations on which all philological research will depend. Rising scholars such as David Smith, David Mimno, Ryan Gabbard, and Gabriel Weaver, originally trained as classicists, were unable to conduct work in machine translation, text mining and general natural language processing that is foundational for classical studies. We may not be able to imagine the shape that our field will assume in the centuries to come, but future change does not absolve us from the obligation to understand what is already possible. None of the PhD programs with which we are familiar has addressed the challenge of producing and supporting those scholars who can show us how to pursue the ancient goals of our field in the rapidly shifting technological spaces within which we live. 122

Third, we need new institutions to provide access to the results of our work. Neither the libraries nor the publishers of the early twenty-first century serve the needs that emerge in this collection. While libraries may survive and indeed flourish as an institution, they will do so by subsuming and transforming the functions that we entrusted to publishers in print culture. We need a small number of library-publishers that can help classicists produce new content and then maintain that content over time. And that content must include not only relatively static documents but, at least for now, a minimal set of executable code: every discipline will probably need at least some services that only experts in the field can create and that are part of the field's core infrastructure. Morphological analysis and lemmatization, mentioned above, are fundamental processes that should be applied automatically to every digital word of Greek and Latin. Classicists may need to develop these systems, but the systems, once developed, need to be preserved as active 123 services along with the XML texts, 2d images, GIS datasets and stable collections.

The seeds of these new organizations are visible in the Digital Knowledge Center in the Johns Hopkins University Library system and the California Digital Library, but we do not yet see in operation a mature model that can serve our needs in the present and expand over time. The Perseus Digital Library thus still finds itself compelled to maintain its own servers as best it can, maintaining services that were innovative a decade ago but are still beyond the capacity of any systems with which we are familiar. Google is moving very quickly in this vacuum. The academic library system failed to address the legal, technical, and financial challenges of converting its retrospective print holdings into digital form. Google Books is rapidly filling the vacuum of collections and services that libraries left. Perhaps it was impossible for our library system, rich in the aggregate, to organize itself. If so, libraries may evolve into a handful of repositories, acting as wholesalers to provide the content by which the Googles, Microsofts, Yahoos and their brethren support the intellectual life of humanity. If the commercial world can generate revenue by providing access to content that anyone can download, then the market may function well enough to provide universal access.[69] 124

Conclusion: Blood for the Shades

> “I see here the spirit of my dead mother; she sits in silence near the blood, and does not look upon the face of her own son or speak to him. Tell me, prince, how she may recognize that I am he.”
>
> [145] So I spoke, and he straightway answered, and said: “Easy is the word that I shall say and put in your mind. Whomsoever of those that are dead and gone you shall allow to draw near the blood, he will tell you true things; but whoever you refuse, he surely will go back again.”
>
> [150] So saying the spirit of the prince, Teiresias, went back into the house of Hades, when he had declared his prophecies; but I remained there steadfastly until my mother came up and drank the dark blood. At once then she knew me. (Odysseus and Teiresias, Homer *Iliad* 11.145-153, after A. T. Murray)
>
> “Fifty sons I had, when the sons of the Achaeans came; nineteen were born to me of the self-same womb, and the others women of the palace bore. Of these, many as they were, furious Ares hath loosed the knees, and he that alone was left me, that by himself guarded the city and the men, him you slew, just now as he fought for his country, even Hector. For his sake have I now come to the ships of the Achaeans to win him back from you, and I bear with me ransom past counting. Nay, have awe of the gods, Achilles, and take pity on me, remembering your own father. See, I am more piteous far than he, and have endured what no other mortal on the face of earth hath yet endured, to reach forth my hand to the face of him that hath slain my sons.”
>
> So he spoke, and in Achilles he roused desire to weep for his father; and he took the old man by the hand, and gently put him from him. So the two thought of their dead, and wept. (Priam and Achilles, Homer *Iliad* 24.495-507, after A. T. Murray)

A new, digital infrastructure provides the explicit subject for this collection of essays. We can create now collections that are larger than any Ptolemy or Cleopatra could have imagined for their Alexandria. We have ever more sophisticated services that can analyze and combine these collections in new ways and even to generate the stuff of new knowledge. And the material systems on which these services are based simply did not exist half a century ago and cost 100,000 times less now than they did a quarter century ago.[70] And if the essays published here have focused upon what we can learn from our textual record, these collections capture sound, images, and data that human hands alone can never transcribe. Indeed, the writing on inscriptions, papyri, and manuscripts now appear as images, open for humans to read and machines to analyze, ready to reveal long forgotten aspects of the living world that produced them. 125

But if everything that we use as a tool is different, nothing that we truly value is new. Like Odysseus in the Underworld, we bring blood to the shades and seek, insofar as possible, to let those who have gone before us to converse with us in their own words. All of us who have studied literature in the academy understand that we can never fully understand our subjects — the very notion of understanding implies a fixity that does not suit complex human beings, filled with 126 contradictory impulses and defined as much by their changing potential for actions as by anything they have done in the past. Priam and Achilles communicate in a single language and understand the cultural backgrounds from which each comes but each of them crosses a gulf as great as that which any mere quantity of time and space can pose. Their moment together has no material effect upon the great events around them. Each will soon suffer a violent death. Troy will fall and a massacre will follow. But the moment above has been powerful for many audiences over the course of almost three millennia, perhaps all the more powerful for the violence that surrounds it.

The future of the past has never been brighter. The digital medium offers new methods with which to make Greco-Roman culture and classical Greek and Latin physically and intellectually accessible to audiences vastly larger and more diverse than was ever feasible in print. The culture of the Greco-Roman world and the languages of classical Greek and Latin can play a fuller role in the cultural memory of all mankind than ever before. The ideas and actions of those who lived in the Greco-Roman world and expressed themselves in Greek and Latin can begin to quicken hearts and fire minds that dream in Chinese, Hindi, and, in the end, every language of humanity. 127

Each of us brings to bear the skills that we have acquired during the time that we have on this earth. Those skills and periods of time vary. Generations pass. Technologies change. Nations rise and fall. Languages fade away and transform themselves beyond recognition. But the memory of classical antiquity has endured over the millennia. All of us who have dedicated our lives to this field — whether we struggle with new technologies or contemplate the record of the past in more traditional ways — are privileged in the subject that we have chosen. We composed these essays in sadness at the loss of our friend, Allen Ross Scaife, but we send them forth in hope as we contemplate the future that Ross helped to make possible. 128

Notes

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

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

[3] George P. Rowell and Company’s American Newspaper Directory. New York: Geo. P. Rowell & Co. 1869. http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0301

[4] See http://www.perseus.tufts.edu/hopper/collection.jsp?collection=Perseus:collection:RichTimes

[5] [Dawkins 1976]. The above is drawn from the Wikipedia entry on the topic: http://en.wikipedia.org/wiki/Meme (accessed August 23, 2008).

[6] For more on the eAQUA project, see their website http://www.eaqua.net/ (accessed October 1, 2008).

[7] [Martin 2006]

[8] http://archimedes2.mpiwg-berlin.mpg.de/archimedes_templates/project.htm; http://archimedes.fas.harvard.edu/

[9] [Marchionini 1994]

[10] [Turing 1950]

[11] [Derrida 1981]

[12] The NSF first began to use Cyberinfrastructure as a strategic term in a 2003 report often called the “Atkins Report”, Atkins, Daniel E., et al. (2003). “Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.” http://www.nsf.gov/od/oci/reports/atkins.pdf (Accessed on October 1, 2008).

[13] For a survey of recent work, see the Cross Language Evaluation Forum, which has been evaluating CLIR since 2000: http://www.clef-campaign.org/

[14] For a discussion of cyberinfrastructure and classical studies, albeit with a focus on scholarly publishing, please see [Pritchard 2008]

[15] For more on CTS, please see [Porter 2006] and [Romanello 2008].

[16] For an overview of the state of the art in this area, please see [Sankar 2006], and for some recent experiments with texts from the Open Content Alliance, please see [Lu 2007].

[17] Some promising initial work in table extraction from digital documents has been reported by [Liu 2007].

[18] For more on Packard's system, see [Packard 1973], and for a discussion of Crane's Morpheus, see [Crane 1991].

[19] The development of the Latin treebank has been documented in [Bamman 2006] and [Bamman 2007].

[20] For some relevant research in this area see [Cardey 2006], and for a recent overview of word sense discovery please see [Pantel 2002].

[21] The entry for the Latin word *ira* ("anger") provides:

Some Words that Regularly Appear with ira					
In Latin Prose (255 total):	indignatio	stimulo	furor	ex-ardeo	saevio
In Latin Poetry (48 total):	suscito	saevio	saevus	Juno	exerceo
In Latin Texts (356 total):	indignatio	stimulo	furor	ex-ardeo	Saturnius

Table 1.

(source)

[22] For an overview of named entity research, see [Nadeau 2007] and for an example of named entity identification in historical texts, see [Byrne 2007].

[23] The named entity system for English has been documented in [Crane 2006a], and its results on one collection evaluated in [Crane 2006e].

[24] For recent work in metrical analysis see [Eder 2007].

[25] For more on word sense disambiguation, please see [Carpuat 2005], and also [Ide 1998].

[26] For some recent overviews of the potential of CLIR for digital libraries and technical issues still to be solved, please see [Jones 2007] and [Petrelli 2006].

[27] For more on plagiarism detection and quotation identification, see [zaslavsky 2001]. Google Books has also recently launched a quotation identification and tracking feature, see [Schilit 2008].

[28] The results of this work can be found in [Ernst-Gerlach 2008].

[29] Recent work in translation detection has been conducted by [Pouliquen 2003].

[30] For recent work in text alignment see [Deng 2006].

[31] For relevant work in this area, please see [Toselli 2007].

[32] Influential work in this area that has greatly informed our own research has been conducted by [Feng 2006].

[33] The literature on the nature of digital editions is quite extensive, for some recent explorations of the topic see [Robinson 2005] and [Dekhtyar 2006].

[34] For further discussion of this issue, see [Monella 2008].

[35] For more on the importance of supporting versioned multitexts, see [Schreibman 2003].

[36] http://wordnet.princeton.edu/; http://www.illc.uva.nl/EuroWordNet/

[37] A bibliography of the research publications using WordNet can be found at http://lit.csci.unt.edu/~wordnet/.

[38] Bamman and Crane in this collection; for sample Treebank data and bibliography, see http://nlp.perseus.tufts.edu/syntax/treebank/

[39] http://www.timeml.org/.

[40] See http://www.getty.edu/research/conducting_research/vocabularies/tgn/; http://stephanus.tlg.uci.edu/canon/fontsel

[41] [Raymond 1987].

[42] See http://heml.mta.ca/heml-cocoon/; http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html

[43] http://cidoc.ics.forth.gr/

[44] [Berti 2007].

[45] For more on this theme see [Campell 2006] and [Pomerantz 2007].

[46] This issue has been discussed by many, including [Smith 2008] and [Johnson 2007].

[47] See http://www.biodiversitylibrary.org/

[48] American Museum of Natural History (New York, NY); The Field Museum (Chicago, IL); Harvard University Botany Libraries (Cambridge, MA); Harvard University, Ernst Mayr Library of the Museum of Comparative Zoology (Cambridge, MA); Marine Biological Laboratory / Woods Hole Oceanographic Institution (Woods Hole, MA); Missouri Botanical Garden (St. Louis, MO); Natural History Museum (London, UK); The New York Botanical Garden (New York, NY); Royal Botanic Gardens, Kew (Richmond, UK); Smithsonian Institution Libraries (Washington, DC).

[49] [Mobley 1998] and [Panitch 2005]. For other on-line sources on the serials crisis, see [Parrott 2004]. The phrase “serials crisis” is sufficiently well-established that it has spawned a Wikipedia entry (http://en.wikipedia.org/wiki/Serials_crisis).

[50] Crane first heard a report about the serials crisis and a presentation arguing that publishers were gouging the market during a meeting held by the Harvard library in 1988.

[51] At http://www.library.cornell.edu/nihmandate/index.html, Cornell University Library posted this summary of the NIH mandate: “Recipients of funding from the National Institutes of Health (NIH) should be aware of a new reporting requirement (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-033.html) that went into effect on April 7, 2008. Principal investigators must ensure that electronic versions of any peer-reviewed manuscripts arising from NIH funding and accepted for publication after that date are deposited in PubMed Central (PMC), NIH's digital archive of biomedical and life sciences journal literature. Full text of the articles will then be made freely available to the public no later than 12 months after publication. The requirement applies to any NIH direct funding, including grants, contracts, training grants, subcontracts, etc. In addition, beginning May 25, 2008, anyone submitting an application, proposal, or progress report to NIH must include the PMC or NIH Manuscript Submission Reference Number when citing applicable articles that arise from their NIH-funded research.”

[52] Harvard University News Release, February 12, 2008: http://www.fas.harvard.edu/home/news-and-notices/news/press-releases/release-archive/releases-2008/scholarly-02122008.shtml.

[53] See http://books.google.com/

[54] http://www.lib.umich.edu/mdp/: accessed August 16, 2008; http://www.lib.umich.edu/libinfo/stats.html: accessed August 16, 2008.

[55] “Searching Free to the Public: Google agrees that to the extent that it or its successors make Digitized Available Content searchable via the Internet, it shall provide an interface for both searching and a display of search results that shall have no direct cost to end users. Violations of this subsection, 4.3, not cured within thirty days of notification by U of M shall terminate U of M's obligations under section 4.4.”: “COOPERATIVE AGREEMENT between Google and the Regents of the University of Michigan/University Library.” Retrieved 8/16, 2008, from http://www.lib.umich.edu/mdp/umgooglecooperativeagreement.html

[56] Google’s approach to mass digitization and open access has been both critiqued and defended for recent examinations of some of the major issues see: [Grafton 2007] and [Kaufman 2007].

[57] http://chs.harvard.edu/, accessed September 30, 2008.

[58] For example, we have started to have a number of early editions of Thucydides scanned and made available such as this edition from 1735 http://www.archive.org/details/tucidideistorico00thucuoft

[59] For a look at how institutional repositories could help to create new models for journals and peer review systems, see [Bankier 2008]; for a good discussion of the need to reevaluate all aspects of scholarly publishing see [Hahn 2008].

[60] The Internet Systems Consortium published for January 2008 an estimate of c. 600,000,000 machines: https://www.isc.org/, accessed August 16, 2008.

[61] For more on the growing recognition of the importance of institutional repositories, see [Hockx-Yu 2006].

[62] ARL Statistics 2005-06: http://www.arl.org/bm~doc/arlstats06.pdf, accessed September 30, 2008.

[63] [Marchionini 2000].

[64] http://epidoc.sourceforge.net/; see also Hugh Cayless' piece in this collection.

[65] LOCKSS stands for "Lots of Copies Keeps Stuff Safe" a program "based at Stanford University Libraries, is an international community initiative that provides libraries with digital preservation tools and support so that they can easily and inexpensively collect and preserve their own copies of authorized e-content." Retrieved from http://www.lockss.org/lockss/Home/.

[66] See [Rosenzweig 2008] (also available at http://chnm.gmu.edu/resources/essays/d/42) and [Giles 2005].

[67] See http://wikimediafoundation.org/wiki/Budget/2005 and http://wikimediafoundation.org/wiki/Planned_Spending_Distribution_2007-2008.

[68] [Bamman 2008].

[69] In the public information Google reports (http://books.google.com/googlebooks/history.html, accessed August 31, 2008): "As part of this fact-finding mission, Larry Page reaches out to the University of Michigan, his alma mater and a pioneer in library digitization efforts including JSTOR and Making of America. When he learns that the current estimate for scanning the university library's seven million volumes is 1,000 years, he tells university president Mary Sue Coleman he believes Google can help make it happen in six." If the source for this figure had imagined the ARL libraries alone dedicating 1% of the $1,000,000,000 collections budget into digital conversion, the $10,000,000 would pay for roughly 300,000 books per year or roughly 16 years for 5,000,000 volumes with the Open Content Alliance Workflow. The library community simply did not think that its retrospective collections were worth the technical, political, and legislative trouble. It will be interesting to see how many observers, a generation from now, will view the leadership of the early twenty first century libraries with sympathy, much less admiration.

[70] In 1982, the Harvard Classics Department paid $34,000 for a 660 megabyte disk ($51 per megabyte). In October 2008, 1 terabyte drives — with more than 1,000 times the capacity of the 600 megabyte drive from 1982 — are available for under $500 (c. $.0005 per megabyte).

Works Cited

Atkins 2003 Warning: Biblio formatting not applied. Daniel E.Atkins. *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. *The Atkins Report*. Arlington. National Science Foundation. 2003. http://www.nsf.gov/od/oci/reports/atkins.pdf.

Bamman 2006 Bamman, David, and Gregory Crane. "The Design and Use of a Latin Dependency Treebank". Presented at *TLT 2006*. *Proceedings of the Fifth International Treebanks and Linguistics Theories Conference* (2006), pp. 67-78.

Bamman 2007 Bamman, David, and Gregory Crane. "The Latin Dependency Treebank in a Cultural Heritage Digital Library". Presented at *LaTeCH 2007*. *Proceedings of the Workshop on Language Technology for Cultural Heritage Data* (2007), pp. 33-40. http://dl.tufts.edu/view_pdf.jsp?pid=tufts:PB.001.002.00002.

Bamman 2008 Bamman, David, and Gregory Crane. "Building a Dynamic Lexicon from a Digital Library". Presented at *JCDL 2008*. *Proceedings of the Eighth ACM/IEE-CS Joint Conference on Digital Libraries* (2008), pp. 11-20.

Bankier 2008 Bankier, Jean-Gabriel, and Irene Perciali. "The Institutional Repository Rediscovered: What can a University do for Open Access Publishing?". *Serials Review* 34 (2008), pp. 21-26.

Berti 2007 BertiM., and V. Costa. "Alexandria and the Mirage of a Million Book Library.". Presented at *The World's Greatest Libraries: From Ancient Alexandria to the 21st Century.* (November 2007).

Byrne 2007 Byrne, Kate. "Named Entity Recognition in Historical Archive Text". Presented at *ICSC 2007*. *Proceedings of the International Conference on Semantic Computing* (2007), pp. 589-596.

Campell 2006 Campbell, Jerry D. "Changing a Cultural Icon: The Academic Library as a Virtual Destination". *Educause Review* 41: 1 (2006), pp. 16-31. http://net.educause.edu/ir/library/pdf/erm0610.pdf.

Cardey 2006 Cardey, Sylvaine, Rosita Chan and Peter Greenfield. "The Development of a Multilingual Collocation Dictionary". Presented at *MLRI 2006*. *Proceedings of the Workshop on Multilingual Language Resources and*

Interoperability (2006), pp. 32-39. http://www.aclweb.org/anthology-new/W/W06/W06-1005.pdf.

Carpuat 2005 Carpuat, Marine, and Dekai Wu. “Word Sense Disambiguation vs. Statistical Machine Translation”. Presented at *ACL 2005*. *Proceedings of the Forty-Third Annual Meeting on Association for Computational Linguistics* (2005), pp. 387-394.

Crane 1991 Crane, Gregory. “Generating and Parsing Classical Greek”. *Literary and Linguistic Computing* 6: 4 (1991), pp. 243-245.

Crane 2006a Crane, Gregory, and Alison Jones. *The Perseus American Collection 1.0*. 2006. http://dl.tufts.edu//view_pdf.jsp?urn=tufts:facpubs:gcrane-2006.00001.

Crane 2006e Crane, Gregory, and Alison Jones. “The Challenge of Virginia Banks: an Evaluation of Named Entity Analysis in a 19th Century Newspaper Collection”. Presented at *JDCL 2006*. *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital Libraries* (2006), pp. 31-40.

Dawkins 1976 Dawkins, Richard. *The Selfish Gene*. Oxford: Oxford University Press, 1976.

Dekhtyar 2006 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 of Digital Libraries* 6: 1 (2006), pp. 55-69.

Deng 2006 Deng, Yonggang, Shankar Kumar and William Byrne. “Segmentation and Alignment of Parallel Text for Statistical Machine Translation”. *Natural Language Engineering* 12: 4 (2006), pp. 1-26.

Derrida 1981 Derrida, Jacques. “Plato's Pharmacy”. In Jacques Derrida, *Dissemination*. Chicago: Chicago University Press, 1981. pp. 61-84.

Eder 2007 Eder, Maciej. “How Rhythmical is Hexameter: A Statistical Approach to Ancient Epic Poetry”. Presented at *Digital Humanities 2007*. (2007). http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=137.

Emerson 1971 Emerson, Ralph Waldo. “The American Scholar”. In Ralph Waldo Emerson, *The Collected Works of Ralph Waldo Emerson, Volume 1: Nature, Addresses, and Lectures*. Cambridge, MA: Harvard University Press, 1971. http://www.apstudent.com/ushistory/docs1801/amrschol.htm.

Ernst-Gerlach 2008 Ernst-Gerlach, Andrea, and Gregory Crane. “Identifying Quotations in Reference Works and Primary Materials”. Presented at *ECDL 2008*. *Proceedings of the Twelfth European Conference on Research and Advanced Technology for Digital Libraries* (2008), pp. 78-87.

Feng 2006 Feng, Shaolei, and R. Manmatha. “A Hierarchical, HMM-based Automatic Evaluation of OCR Accuracy for a Digital Library of Books”. Presented at *JCDL 2006*. *Proceedings of the Sixth ACM/IEEE-CS Joint Conference on Digital Libraries* (2006), pp. 109-118.

Foxe 1965 FoxeJohn. *Acts and Monuments, Volume 5*. Arno Press, 1965.

Giles 2005 Giles, Jim. “Internet Encyclopaedias go Head to Head”. *Nature* 438 (2005), pp. 900-901.

Grafton 2007 Grafton, Anthony. “Future Reading: Digitization and Its Discontents”. *The New Yorker* (November 5, 2007). http://www.newyorker.com/reporting/2007/11/05/071105fa_fact_grafton?currentPage=all.

Hahn 2008 Hahn, Karla L. “Talk About Talking About New Models of Scholarly Communication”. *Journal of Electronic Publishing* 11: 1 (2008). http://hdl.handle.net/2027/spo.3336451.0011.108.

Hockx-Yu 2006 Hockx-Yu, Helen. “Digital Preservation in the Context of Institutional Repositories”. *Program: Electronic Library and Information Systems* 40: 3 (2006), pp. 232-243.

Ide 1998 Ide, N., and J. Veronis. “Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art”. *Computational Linguistics* 24: 1 (1998), pp. 1-40.

Johnson 2007 JohnsonR. K. “In Google's Broad Wake: Taking Responsibility for Shaping the Global Digital Library”. *ARL: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC* 250 (February 2007), pp. 1-17. http://www.arl.org/storage/documents/publications/arl-br-250.pdf.

Jones 2007 Jones, Gareth J. F., Ying Zhang, Eamonn Newman, Fabio Fantino and Franca Debole. “Multilingual Search for Cultural Heritage Archives via Combining Multiple Translation Resources”. Presented at *LaTeCH 2007*. *Proceedings of the Workshop on Language Technology for Cultural Heritage Data* (2007), pp. 81-88. http://www.aclweb.org/anthology-new/W/W07/W07-0911.pdf.

Kaufman 2007 Kaufman, Peter B., and Jeff Ubois. “Good Terms: Improving Commercial-Noncommercial Partnerships for Mass Digitization”. *D-Lib Magazine* 13: 11/12 (2007). http://www.dlib.org/dlib/november07/kaufman/11kaufman.html.

Liu 2007 Liu, Ying, Kun Bai, Prasentjit Mitra and C. Lee Giles. “TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries”. Presented at *JCDL 2007*. *Proceedings of the JCDL 2007* (2007), pp. 91-100.

Lu 2007 Lu, Xiaonan, James Z. Wang and C. Lee Giles. “Intelligent Parsing of Scanned Volumes for Web Based Archives”. Presented at *ICSC 2007*. *Proceedings of the International Conference on Semantic Computing* (2007), pp. 559-568.

Marchionini 1994 Marchionini, Gary, and Gregory Crane. “Evaluating Hypermedia and Learning: Methods and Results from the Perseus Project”. *ACM Transactions on Information Systems* 2: 1 (1994), pp. 5-34.

Marchionini 2000 MarchioniniGary, and Ross Scaife, et al. *1997-2000 Final Evaluation Report on the Perseus Project Publication Model*. 2000. http://ils.unc.edu/~march/perseus/final_report.pdf.

Martin 2006 Martin, Thomas R. *Ancient Greece: From Prehistoric to Hellenistic Times*. New Haven: Yale University Press, 1996.

McCown 2005 McCown, Frank, Sheffan Chan, Michael L. Nelson and Johan Bollen. “The Availability and Persistence of Web References in D-Lib Magazine”. Presented at *IWAW 2005*. *Proceedings of the Fifth International Web Archiving Workshop and Digital Preservation* (2005).

Mobley 1998 MobleyEmily R. “Ruminations on the Sci-Tech Serials Crisis.”. *Issues in Science and Technology* 20 (Fall 1998). http://www.istl.org/98-fall/article4.html.

Monella 2008 Monella, Paolo. “Towards a Digital Model to Edit the Different Paratextual Levels With a Textual Tradition”. *Digital Medievalist* 4 (2008). http://www.digitalmedievalist.org/journal/4/monella/.

Nadeau 2007 Nadeau, David, and Satoshi Sekine. “A Survey of Named Entity Recognition and Classification”. *Lingusticae Investigationes* 30: 1 (2007), pp. 3-26.

Packard 1973 Packard, David W. “Computer-Assisted Morphological Analysis of Ancient Greek”. Presented at *COLING 1973*. *Proceedings of the Fifth Conference on Computational Linguistics* (1973), pp. 343-355.

Panitch 2005 PanitchJ. M., and S. Michalak. *The Serials Crisis*. *UNC-Chapel Hill Scholarly Communications Convocation*. 2005. http://www.unc.edu/scholcomdig/whitepapers/panitch-michalak.html.

Pantel 2002 Pantel, Patrick, and Dekang Lin. “Discovering Word Senses from Text”. Presented at *ACM SIGKDD 2002*. *Proceedings of the Eighth ACM Special Interest Group on Knowledge Discovery and Data Mining* (2002), pp. 613-619.

Parrott 2004 Parrott, Jim, ed. *The Crisis in Scholarly Publishing*. University of Waterloo Library, 2004. http://www.lib.uwaterloo.ca/society/crisis.html.

Petrelli 2006 Petrelli, Daniela, Stephen Levin, Mark Sanderson and Micheline Sanderson. “Which User Interaction for Cross-Language Information Retrieval? Design Issues and Reflections”. *Journal of the American Society for Information Science and Technology* 57: 5 (2006), pp. 709-722.

Pomerantz 2007 Pomerantz, Jeffrey, and Gary Marchionini. “The Digital Library as Place”. *Journal of Documentation* 60: 4 (2007), pp. 505-533.

Porter 2006 Porter, Dorothy, William Du Casse, Jerzy Jaromczyk, Neal Moore, Ross Scaife and Jack Mitchell. “Creating CTS Collections”. *Digital Humanities 2006* (2006), pp. 269-274. http://www.csdl.tamu.edu/~furuta/courses/06c_689dh/dh06readings/DH06-269-274.pdf.

Pouliquen 2003 Pouliquen, Bruno, Ralf Steinberger and Camelia Ignat. “Automatic Identification of Document Translations in Large Multilingual Document Collections”. Presented at *RANLP 2003*. *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (2003), pp. 401-408.

Pritchard 2008 Pritchard, David. “Working Papers, Open Access, and Cyber-infrastructures in Classical Studies”. *Literary and Linguistic Computing* 23: 2 (2008), pp. 149-162. http://ses.library.usyd.edu.au/handle/2123/2226.

Raymond 1987 Raymond, Darrell R., and Frank W. Tompa. “Hypertext and the new Oxford English Dictionary”. Presented at *HYPERTEXT 1987*. *Proceedings of the ACM Conference on Hypertext* (1987), pp. 143-153.

Robinson 2005 Robinson, Peter M.W. “Current Issues in Making Digital Editions of Medieval Texts -- or, Do Electronic Scholarly Editions Have a Future?”. *Digital Medievalist* 1: 1 (2005). http://www.digitalmedievalist.org/journal/1.1/robinson/.

Romanello 2008 Romanello, Matteo. “A Semantic Linking Framework to Provide Critical Value-Added Services for E-Journals on Classics”. Presented at *ELPUB 2008* (2008). *Proceedings of the Twelfth International Conference on Electronic Publishing*. http://elpub.scix.net/cgi-bin/works/Show?401_elpub2008.

Rosenzweig 2008 Rosenzweig, Roy. “Can History be Open Source: Wikipedia and the Future of the Past?”. *Journal of American History* 93: 1 (2006), pp. 117-146. http://chnm.gmu.edu/resources/essays/d/42.

Sankar 2006 Sankar, K. Pramod, Vamshi Ambati, Lakshmi Pratha and C.V. Jawahar. “Digitizing a Million Books: Challenges for Document Analysis”. *Document Analysis Systems VII* (2006), pp. 425-436.

Schilit 2008 Schilit, Bill N., and Okan Kolak. “Exploring a Digital Library through Key Ideas”. Presented at *JCDL 2008*. *Proceedings of the Eighth ACM/IEEE-CS Joint Conference on Digital Libraries* (2008), pp. 177-186.

Schreibman 2003 Schreibman, Susan, Amit Kumar and Jarom McDonald. “The Versioning Machine”. *Literary and Linguistic Computing* 18: 1 (2003), pp. 101-107.

Shirky 2008 Shirky, Clay. *Here Comes Everybody: The Power of Organizing Without Organizations*. New York: Penguin Press, 2008.

Smith 2008 Smith, Abby. “The Research Library in the 21st Century: Collecting, Preserving, and Making Accessible Resources for Scholarship”. *Council on Library and Information Resources (CLIR)* 142 (2008), pp. 13-20. http://www.clir.org/pubs/reports/pub142/smith.html.

Toselli 2007 Toselli, Alejandro, Veronica Romero and Enrique Vidal. “Viterbi Based Alignment Between Text Images and Their Transcripts”. Presented at *LaTeCH 2007*. *Proceedings of the Workshop on Language Technology for Cultural Heritage Data* (2007), pp. 9-16.

Turing 1950 Turing, Alan. “Computing Machinery and Intelligence”. *Mind* 59: 236 (1950), pp. 433-460.

Von Ranke 1973 Von Ranke, Leopold. *The Theory and Practice of History*. Edited by Georg G. Iggers and Konrad Von Moltke. New York: Irvington Publishers, 1973.

von Humboldt 1821 Wilhelm von Humboldt, “Lecture to the Prussian Academy,” 1821. From a lecture delivered to the Prussian Academy of Sciences in 1821, quoted in [Von Ranke 1973, 21].

zaslavsky 2001 Zaslavsky, Arkady B., Alejandro Bia and Krisztián Monostori. “Using Copy-Detection and Text Comparison Algorithms for Cross-Referencing Multiple Editions of Literary Works”. Presented at *ECDL 2001*. *Proceedings of the Fifth European Conference on Research and Advanced Technology for Digital Libraries* (2001), pp. 103-114.