

ePhilology: when the books talk to their readers¹

Gregory Crane
David Bamman
Alison Babeu
Perseus Project
Tufts University

[Forthcoming in Blackwell Companion to Digital Literary Studies, Ray Siemens and Susan Schreibman eds. (2007)]

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.

Plato, *Phaedrus* 275d

Introduction

This paper suggests directions in which an ePhilology may evolve. Philology here implies that language and literature are the objects of study but assumes that language and literature must draw upon the full cultural context and thus sees in philological analysis a starting point for the *scientia totius antiquitatis* – the systematic study of all ancient culture. The term ePhilology implicitly states that, while our strategic goal may remain the *scientia totius antiquitatis*, the practices whereby we pursue this strategic goal must evolve into something qualitatively different from the practices of the past.

Digital technology is hardly new in classics: there are full professors today who have always searched large bodies of Greek and Latin, composed their ideas in an electronic form, found secondary sources on-line and opportunistically exploited whatever digital tools served their purposes.² Nevertheless, the inertia of prior practice has preserved intact the forms that evolved to exploit the strengths and minimize the weaknesses of print culture: we create documents that slavishly mimic their print predecessors; we send these documents to the same kinds of journals and publishers;³ our reference works

¹ The work described here builds on support from a variety of sources, including the Digital Library Initiative, Phase 2, the National Endowment for the Humanities, the National Science Foundation, and the Institute for Museum and Library Services. Many individuals have contributed. We mention in particular Carla Brodley, Lisa Cerrato, David Mimno, Adrian Packel, D. Sculley, and Gabriel Weaver.

² For some reviews of how technology has been used within classics, please see: (Crane 2004), (McManus 2003), (Latousek 2001), and (Hardwick 2000)

³ Classicists were quick to embrace the Bryn Mawr Classical Review (<http://ccat.sas.upenn.edu/bmcr/>), which began publication in 1990 as a mailing list. BMCR was successful for three reasons: first, it used e-mail to speed up the pace of scholarly communication, thus addressing a single, nagging problem; second, the electronic form allowed BMCR greater flexibility than its print counterparts, allowing it to accept a greater range of reviews, thus encouraging a wider range of submissions; third, its articles were, and

and editions have already begun to drift out of date before they are published and stagnate thereafter; even when new, our publications are static and cannot adapt themselves to the needs of their varying users; while a growing, global audience could now find the results of our work, we embed our ideas in specialized language and behind subscription barriers which perpetuate into the twenty-first century the miniscule audiences of the twentieth.⁴

This paper makes two fundamental arguments. First, it assumes that the first generation of digital technology has only laid the groundwork for substantive change in classics and the humanities. Second, it advances arguments about what form an optimal digital future should assume. While Greek and Latin provide the focus for this paper, the arguments apply in various ways to many areas within the humanities.

At least six features distinguish emerging digital resources: (1) they can be delivered to any point on the earth and at any time; (2) they can be fundamentally hypertextual, supporting comprehensive links between assertions and their evidence; (3) they dynamically recombine small, well defined units of information to serve particular people at particular times; (4) they learn on their own and apply as many automated processes as possible, not only automatic indexing but morphological and syntactic analysis, named entity recognition, knowledge extraction, machine translation etc., with changes in automatically generated results tracked over time; (5) they learn from their human readers and can make effective use of contributions, explicit and implicit, from a range of users in real time; (6) they automatically adapt themselves to the general background and current purposes of their users.

Print culture gave us expensive distribution by which we could send static documents to a few thousand restricted locations. If we can deliver information to any point on the earth and we can tailor that information to varying backgrounds and immediate purposes of many people, we can thus address audiences far beyond the physical and, indeed, cultural limitations which communication – oral and print – has imposed.

In the *Phaedrus*, Plato's Socrates, a fictive rendering of a historical character scratched into life by pen and preserved as a pattern of ink, critiques writing – and thus the very medium in which he exerts a living presence to this day. A generation ago, Derrida famously expanded upon the observation that writing is not so much a cure as a poison for memory⁵ – in a look-up culture not only do our memories decay but we lose in some measure that instant and deep recognition which integrated knowledge alone can spark. The critique in the *Phaedrus* is profound and addresses all technologies which represent information abstracted not only from the brain but also from the personal context in which much learning occurs. Plato's arguments have been echoed ever since, consciously or not – many of us in the first generation of a television society heard

remain, electronic analogues of print: they do not challenge their authors to rethink the substantive form of their work. The Stoa publishing consortium, by contrast, began in 1997 and has supported a range of more innovative projects (including the Demos project described below).

⁴ For some recent overviews of the issues with scholarly publishing, please see (Unsworth 2003).

⁵ (Derrida 1972).

similar criticisms from our parents and, in turn, directed these to our net-oriented children.⁶

The quote that begins this essay, however, directs a criticism which is just as trenchant but has attracted less attention. All products of information technology – paintings and poems, novels and newspapers, movies and music – have been static since our ancestors first scratched diagrams in the dirt or pressed visions of their world on the walls of caves. Other human hands could add or destroy, but the products of our hands could do nothing but decay, prey to the scorching sun, the worm or the slow fires of acid within. We can direct our questions to the written word or to the most lifelike painting, but we can expect only silence.

Now, however, we have created cultural products that can respond, systems that can change and adapt themselves to our needs. Millions of people around the world will, on the day that I compose these words, seek directions from a mapping service. Natural language, mathematical formulae, and visual representations of space will interact to generate tailored itineraries, with estimates of time and customizable maps illustrating the journey from one point to another, in some cases speaking their directions in an expanding suite of languages. We should not confuse the humble and well-defined goals of such tools with their significance in the evolution of humanity and indeed life.

The great question that we face is not what we can do but what we want to accomplish. The tools at our disposal today, primitive as they may appear in the future, are already adequate to create a dynamic space for intellectual life as different from what precedes it as oral culture differs from a world of writing. At one level, little will change – the Homeric epics, products of an oral culture ironically preserved in writing, are arguably as successful as cultural products as anything which followed: the ceiling of human creativity has not changed in three thousand years of increasingly sophisticated information technology – an observation that we should consider as we fret over the codex and print.⁷

Nevertheless, we can now plan for a world where ideas cross from language to language and from culture to culture with a speed and authenticity far beyond what we have ever experienced. Consider curious minds in Beijing or Damascus a generation from now who encounter something that sparks their interest in the Greco-Roman world. It could be a film or a popular novel translated into Chinese or Arabic or a game that carries them through a virtual space. It could even be something their formal education which, as occasionally happens, fires their imagination. The internet as we have it has already increased the chances for such encounters and provided unprecedented opportunities for Beijing and Damascus to learn about ancient Greece and Rome or other cultures.

We can, however, do more. The intellectually alive mind asks about a Greek author, perhaps a widely translated one (such as Homer) or perhaps not. Background

⁶ For a discussion of fears that Google and the digitization of libraries will lead to serious decontextualization of learning, see (Garrett 2006).

⁷ Dino Franco Felluga discusses this issue as well in regards to literary studies; please see (Felluga 2005).

information and the text itself are translated into the Chinese or Arabic. The inquirer has developed a profile, not unlike her medical history, which can record the classes she has taken, the books she has read, the movies she has seen, the games she has played, and the questions that she has posed.⁸ The personal reading agent can compare this profile, eagerly developed and shared only in part and under strict conditions, against the cultural referents implicit in the author or text of interest, then produce not only translations but personalized briefing materials – maps, timelines, diagrams, simulations, glossary entries – to help that reader contextualize what she has encountered. As the reader begins to ask questions, the system refines its initial hypotheses, quickly adapting itself to her needs.⁹ As the system changes, it inspires new kinds of inquiry in the reader, creating a feedback loop that encourages their conversation to evolve. Far from the static and one-sided interaction of Plato’s complaint, this is the definition of dialectic.

As this paper will suggest, we already possess the technology to build a system of this type that will be effective in many cases: the professional classicist moving into early modern Latin or even tracking developments in his or her own field, with text mining identifying trends in the secondary literature or phenomena in the source texts.¹⁰

The question that we face is much deeper than the challenge of producing more or, preferably, better articles and monographs. We must more generally ask what kind of space we wish to produce in which to explore the linguistic record of humanity – whether we are contemplating the *Odyssey*, administrative records from Sumer, or tracing mathematical thought through Greek and Arabic sources. More important perhaps than the question of what we can do may be the opportunity to redefine who can do what – to open up intellectual life more broadly than ever before and to create a fertile soil in which humanity can cultivate the life of the mind with greater vigor and joy.

Background

The systematic application of computing technology to classical languages began in 1968, when David Packard had toiled with primitive computing in the basement of Harvard’s Science Center to produce a full concordance to Livy. The resulting massive print volumes were both a fundamental new tool and a staple at Harvard University Press remainder sales of the 1970s, illustrating both the potential of even simple electronic tools and the limitations of the codex. Three fundamental developments quickly followed.

⁸ The idea of a permanent personal digital archive or storehouse of lifetime memories and knowledge has been well articulated by the creators of MyLifeBits (Gemmell 2006). Neil Beagrie has also explored this concept (Beagrie 2005).

⁹ A wealth of research has been conducted into how systems can best automatically adapt themselves to the needs of different readers, such as (Russell 2003), (Dolog 2004), (Niederee 2004), (Rouane 2003), (Wang 2004), and (Terras 2005).

¹⁰ Text mining is increasingly being used in humanities applications; see, for example, (Kirschenbaum 2006) and (Xiang 2006).

First, The *Thesaurus Linguae Graecae*,¹¹ founded in 1972, began developing what would be called a digital library of classical Greek literature. A third of a century later, the TLG has completed its initial goal of digitizing all published Greek literature up through 600 AD and has extended its coverage through the Byzantine period and beyond.¹² The TLG thus provided the first digital well-curated collection of digital resources in classics.

Second, David Packard began in the 1970s to develop a system not only to work with collections such as the TLG but to provide the first computerized typesetting and word processing for Greek.¹³ At the Boston APA convention of 1979, for example, Packard could show a working Ibycus computer system. Based on a Hewlett Packard Minicomputer, the Ibycus included a unique operating system designed for classics. The Ibycus was, by the standards of the early twenty-first century, astonishingly expensive – it cost tens of thousands of dollars – but it provided scholars with services they needed not only to exploit the TLG but to write and publish. Its contributions were so important that more than a dozen departments raised the necessary capital.

Third, the TLG and Ibycus System were the products of two distinct organizations, thus promoting a separation of data from service providers and opening the way for a range of entrepreneurs to create additional services and solutions.¹⁴ The TLG website lists more than a dozen packages that were developed to work with the CD ROM texts.

A generation later, classicists still depend upon texts and services designed in the 1970s. Figure 1 illustrates the results from a sample search of the TLG in May 2006 as suggested on the TLG website. The system reflects decades of investment, both from subscriptions and from grants (e.g., a 2000 \$235,000 grant from the National Endowment for the Humanities that provided partial support for “restructuring of data and development of an online search and retrieval system for the *Thesaurus Linguae Graecae*.”¹⁵) The resulting in-house system provides a fast, reliable service on which Hellenists depend, especially since the TLG no longer updates its CD ROM and thus does not generally distribute source texts published after the February 2000 TLG E Disk.¹⁶

It would be hard to overstate the importance of searchable text corpora. Classicists are also fortunate to have access to the Packard Humanities Institute CD ROM for Latin literature, as well as proprietary commercial databases such as the *Biblioteca Teubneriana Latina*.¹⁷ Classicists have become accustomed to scanning wide swathes of Greek and Latin literature, with full professors today who have never known a world without searchable texts. Many take for granted this core infrastructure and, when asked, admit that these tools have had far more impact upon the questions that they ask and the research that they conduct than they readily articulate. An analysis of primary source

¹¹ <http://www.tlg.uci.edu/>.

¹² For one exploration of the impact of the TLG on classical scholarship, please see (Ruhleder 1995).

¹³ For a discussion of some of this work, see (Packard 1973).

¹⁴ Crane 2004.

¹⁵ <http://www.neh.gov/news/awards/preservation2000.html>.

¹⁶ <http://www.tlg.uci.edu/CDROME.html>.

¹⁷ (*Biblioteca Teubneriana Latina* 2004)

citations in the classics journals of JSTOR would give us a better appreciation of the impact which these collections have had upon published scholarship.

TLG Texts
Settings
Greek display?
Unicode
Greek input?
Beta Code
Results per page?
5
Lines of context? 3
Perseus links? No links
Beta escapes? Show
With diacritics?
Ignore diacritics
 Order by date?
 Case sensitive?
 Adscript as subscript?
 Ignore partial words?
Full Corpus Search
Simple
Advanced
TLG Canon
Keyword Search
 Go
Search by
Author
Work
Publication
Site Links

Search Results

Search for ἀποδεξ:

1. [Thucydides Hist., *Historiae*. {0003.001} Book 3 chapter 57 section 1 line 5. \(Browse\)](#)
τήνδε, ἐπαινούμενοι δὲ περὶ οὐδ' ἡμῶν μεμπτῶν), ὁρᾶτε ὅπως
μὴ οὐκ ἀποδέξωνται ἀνδρῶν ἀγαθῶν πέρι αὐτοὺς ἀμείνους (5)
ὄντας ἀC0;ρεπές τι ἐπιγνώναι, οὐδὲ πρὸς ἱεροῖς τοῖς κοινοῖς
2. [Thucydides Hist., *Historiae*. {0003.001} Book 7 chapter 48 section 3 line 4. \(Browse\)](#)
οὐκ ἔφη ἀπάξειν τὴν στρατιάν. εὖ γὰρ εἰδέναι ὅτι Ἀθη-
ναῖοι σφῶν ταῦτα οὐκ ἀποδέξονται, ὥστε μὴ αὐτῶν ψηφισα-
μένων ἀπελθεῖν. καὶ γὰρ οὐ τοὺς αὐτοὺς ψηφιεῖσθαι τε (5)
3. [Euripides Trag., *Helena*. {0006.014} Line 832. \(Browse\)](#)
{Ελ.} ὡς οὐκ ἄχρωστα γόνατ' ἐμῶν ἔξει χερῶν.
{Με.} φέρ', ἦν δὲ δὴ νῶν μὴ ἀποδέξεται λόγους;
{Ελ.} θανῆ· γαμοῦμαι δ' ἡ τάλαιν' ἐγὼ βία.
4. [Plutarchus Biogr. et Phil., *Galba*. {0007.065} Chapter 16 section 1 line 6. \(Browse\)](#)
δεῖπνον (ἀκρόαμα δὲ ἦν ὁ Κάνος εὐδοκμοῦμενον) (5)
ἐπαινέσας καὶ ἀποδεξάμενος ἐκέλευσεν αὐτῷ
κομσθῆναι τὸ γλωσσόκομον· καὶ λαβὼν χρυσοῦς
5. [Isocrates Orat., *In Euthynum*. {0010.001} Section 18 line 1. \(Browse\)](#)

Figure 1: TLG Search, May 17, 2006

In the past thirty years more texts have been added but the essential services and underlying data model visible to the classical community has not changed. The TLG, as it appeared in May 2006, is selected for analysis because it has successfully served, and continues to serve, the field and provides a standard of excellence, in terms of continuity and quality of service, but the analysis offered below applies to many efforts in classics and the humanities. The goal is not to diminish the importance of what TLG and projects like it have contributed but, by describing the state of the art as it existed when this article was written, to suggest future movements for classics and the humanities.

- **String based searching:** As this article is being written, users do not search for a lexeme (e.g., APODEIKNUMI) but for strings which to find inflected forms.¹⁸ This reduces precision (the string above would, for example, locate not only forms of the verb APODEIKNUMI but the noun APODECIS) and recall (to locate all forms of the verb one would need to search for other strings, e.g., APEDEC, APODEIKN, APODEIC, APODEXQ, etc.)

¹⁸ Maria Pantelia, the director of the TLG reported (private communication, September 2006) that lemmatized searching was in active development and would become part of the core TLG functionality.

Users need to find many other patterns and we need research and development on a range of searches.¹⁹ Lemmatized searches allow users to query a dictionary entry and identify all known inflected forms. Collocational analysis allows users to find words that co-occur with unusual frequency and thus to uncover idiomatic expressions.²⁰ Users also need to be able to locate syntactic patterns: e.g., what subjects and objects does a particular verb take? How often does the verb actually take the dative in a given corpus? What adjectives modify particular nouns? They need to search for people and places, identifying not only all Alexanders and Alexandrias but also be able to locate references to the particular Alexander and Alexandria in which they are interested. They should be able to find basic propositional patterns: e.g., at what locations does person X appear in within the corpus?²¹ They should be able to apply intelligent clustering, automatic summarization and text mining to searches that produces thousands of results.²² They should be able to search for secondary sources that talk about directly or are generally relevant to any given passage.

- Texts are encoded as page surrogates: venerable Beta code markup tags the speakers in the Euripides search results pictured above. In the Thucydides and Plutarch results, the electronic text faithfully reproduces the line-breaks (including hyphenization) of the print original. Users cannot exploit semantic markup (e.g., search and compare results from the language of Helen and Menelaus in the *Helen* of Euripides, separate results from spoken vs. narrative text in Thucydides). Even the section breaks are only approximately encoded, with section breaks, for example, simply inserted at the start of the line rather than in their proper position.²³ It is not difficult to convert the page layout Beta encoding of the TLG into TEI compliant SGML or XML,²⁴ but fuller conversion requires substantial editing with enough human interpretation of the meaning implicit in the page layout for a true XML version to appear as a new edition in its own right. The cost of analyzing and formatting a complex document (such as a play) is

¹⁹ There is a growing body of research into the need for more complex linguistic querying capabilities, particularly with historical language materials, please see (de Jong 2005), (Egan 2005), and (Gerlach 2006).

²⁰ See, for example, (Church and Hanks 1989) and (Justeson and Katz 1995).

²¹ The Perseus Digital Library has done extensive research in terms of the importance of named entity recognition and searching; please see (Crane and Jones, 2006), (Smith and Crane, 2001).

²² For an example of a prototype system that supports many of these features, please see (Ignat 2005).

²³ A search for –pemp– turns up “(4.) OI)KH/TORAS A)POPE/MPEIN. OI(DE\ *)EPIDA/MNIOI OU)DE\N AU)TW=N U(P-” with the label Thucydides “Book 1 chapter 26 section 4 line 1.” In fact, the word is part of section 3, with section four beginning in the middle of the print line after the period. Simple programming can capture most of these section breaks, although some lines have more than one full stop and editors may use commas – or nothing – to mark the divisions of established units.

²⁴ In the late 1990s, while Theodore Brunner was director of the TLG, David Smith of the Perseus Project created an SGML version of the TLG that validated against the TEI DTD. Mark Olsen of ARTFL also created a similar experimental version at the University of Chicago. In both cases, understanding the idiosyncratic reference encoding of the TLG proved the major barrier.

comparable to the cost of double keyed professional data entry.²⁵

- Texts represent only a single, isolated edition: After consulting with the scholarly community, the TLG chose to encode only the consolidated text, leaving aside variants and providing only a single edition each author.²⁶ At the time, the added cost and complexity were determining factors. This initially reluctant measure has become policy: the TLG suppressed older editions, removing them from circulation and replacing them completely.²⁷ Rather than letting users search both the Murray (which was on the D Disk) and the Diggle edition (which took its place on the E Disk), users received just the one, more recent edition and (to use the TLG's own language) "suppressed" the older editions.²⁸
- Limited interoperability: The TLG does build in some measure on third party efforts: the TLG can, for example, add links to the open access morphological and lexicographic data at the Perseus Digital Library but there are no clear methods whereby third party systems can interact with the TLG. Even sites that erect subscription barriers around their data do not have to be data silos. The TLG Canon could be distributed, at least in part, via the Open Archive Initiative, a low-barrier approach well suited to distributing cataloguing data.²⁹ This would allow pointers to TLG texts to appear in library catalogues and for third party searching and text mining to add value to the base data. At a more advanced level, even if the TLG does not choose to distribute its newer texts, it could make search results available via an API so that subscribing third parties could efficiently analyze the results of searches and/or create customized front ends. The emerging Classical Texts Services protocol³⁰ would provide a consistent method whereby systems could extract labeled chunks of text – a crucial function as dynamically generated documents emerge.
- Texts cannot be readily repurposed or circulate freely: Publishers assert copyright to the editions which they publish. The legality of this claim is by no means clear,³¹ and publisher claims represent aspiration rather than settled law – Norton

²⁵ The largest Greek dramas are, with extensive XML markup, just over 120,000 bytes and would cost \$120 to \$180 to enter, depending on the vendor.

²⁶ Research into variant editions and how best represent this information digitally has received a growing amount of attention, for example see (Dekhytar 2005), (Pierazzo 2006), (Schmidt 2005), and (Audenaert 2005) and (Riva 2005), and for an example in classics (Bodard 2006).

²⁷ <http://www.tlg.uci.edu/CDEworks.html#supp>.

²⁸ The online TLG does not seem to provide any information about the texts that have been "suppressed," in effect consigning these editions to an electronic *damnatio memoriae*. A print copy of the second edition of the TLG canon preserves the fact that the TLG had originally contained the Murray edition of Euripides. The online TLG canon simply lists the Diggle edition of Euripides now included in the TLG.

²⁹ <http://www.openarchives.org/>

³⁰ (Blackwell and Smith 2005). For some examples of how the CTS protocols are being used, please see (Porter, et. al, 2006)

³¹ According to at least one participant at the international gathering of Hellenists which launched the TLG in the early 1970s, the experts in the field assumed that the texts of ancient authors, as published in editions, were not copyrightable. We need automated methods with which not only to compare but to quantify the differences between various electronic editions of the same text. Preliminary analysis suggests

went so far as to claim copyright to the through-line-numbers in their published facsimile of the First Folio (Hinman 1968) – in fact, a computer program will generate the through line numbers by mechanically counting lines and thus no recognizable “original expression” is in play. Publishers have, however, traditionally charged permission fees for materials that were in the public domain and an exploration of rights and practices would provoke interesting lines of inquiry.³² The threat of legal action, however frivolous, has exercised a chilling effect upon scholarship. The publishing institutions that exist to facilitate the exchange of ideas thus choke the circulation of primary materials, constrain the fundamental moral right of academic authors to reach the broadest possible audience, and restrict scholarly activity.³³ With no new TLG CD ROMs, an emphasis on a single proprietary site, and no interoperability (not even an OAI harvestable version of the TLG Canon), Hellenic studies have, if anything, taken a step backwards.

The limitations described above have been acceptable because they support the practices of print culture. Textual corpora such as the TLG, whether on the Web or on CD ROM, are immense, dynamic, flexible concordances. They thus support traditional work but also provide no incentive for innovative forms of publication. The monolithic web site isolates classicists from the electronic infrastructure which supports them. If our goal is to produce more and better researched articles and monographs – if we think that the answer to the crisis in academic monographs is to produce more content – then the status quo will serve us well.³⁴

The Future in the Present

At this point, we return to the six features that, at least in part, distinguish digital from print publication. While work remains at an early stage of development, progress is being made in all six areas. The following section illustrates these points primarily with work done associated with Perseus for classics, but Perseus and the field of classics are only components of a much larger process.³⁵

that changes from one edition to another are comparable to copy-editing. The best model for editors employed by academic institutions may thus be a work-for-hire, with the rights holders more properly being institutions who paid their salary.

³² The representative of one UK publisher stopped at Perseus years ago en route, as he informed us, to assert rights to electronic versions of texts that a third project had entered. We paid \$7,000 for rights to two editions – only to discover that those editions had unambiguously gone into the public domain by UK law and had never been under copyright in the US. Another US publisher that had knowingly published materials in the public domain reportedly charges permissions fees for these materials for which it has no legal rights.

³³ For more on the issue of the public domain and copyright issues in the face of mass digitization, please see (Thatcher 2006) and (Travis 2005).

³⁴ Classicists define their own conventions of what does and does not count, and we can accept monographs published in emerging institutional repositories – in effect, we would return to a scholarly publication model, separate from university and commercial presses, that has served us well in the past.

³⁵ For further discussion of Perseus examples, please see (Crane et al. 2006).

Global access

Library subscription budgets shield many scholars – especially those at the most prestigious institutions – from the economic realities with which libraries struggle. Many – probably most – do not realize that the scholarly resources – much of it in the public domain – on which they daily rely are available only through expensive subscriptions. Various open access movements have attacked this problem – rarely with support, not infrequently with scorn, from academics: Project Gutenberg began in 1971 (one year before the TLG), hosts a library of 18,000 public domain books and downloads two million of these each month.³⁶ More recently, Google Library and the Open Content Alliance (OCA) have set out to digitize the entire published record of humanity. Each pursues contrasting rights regimes: Google retains its collection for its proprietary use, while the OCA is building an open source collection: Yahoo and Microsoft are both backing OCA, with each planning to provide its own set of unique services to the shared content. Both Google Library and the OCA are, however, open access – the business models of Google, Yahoo and Microsoft all depend upon maximizing their audiences.³⁷ Open access seems to them to be a better engine for revenue generation than subscription models.³⁸

Within classics, the Latin Library dramatized the hunger for open-source primary materials. Frustrated by proprietary text corpora, members of the community, most from outside of academia, have spontaneously digitized almost all classical Latin, and a growing body of post-classical, Latin literature and made it freely accessible at a single site.³⁹ It is easy to criticize this work: original scholarship resides along with texts bear the unnerving label “from an unidentified edition,” while other texts combine multiple editions without substantive documentation.⁴⁰ The site reflects a wide-spread and heartfelt desire to assemble a critical mass of freely accessible, Latin texts. While professional scholars can criticize some of the texts, we should also ask ourselves why the community felt it necessary to do so much work to establish such a basic service. Were the publications that we composed with proprietary databases a greater contribution to intellectual life than a universally accessible library of primary texts?

From the beginning of its Web presence in 1995, Perseus provided open access to all of its holdings not otherwise restricted by third party rights.⁴¹ More recently, members of the community – especially the rising generation of classicists – have argued forcefully that all core materials should be available under an open source license, allowing third parties to repurpose what we have begun. We have thus moved beyond open access and to open source for all materials to which we have rights. We chose a Creative Commons

³⁶ http://www.gutenberg.org/wiki/Main_Page

³⁷ For an extensive discussion of the Google Library project, please see (MacColl 2006), for the Open Content Alliance (Tennant 2005).

³⁸ For a comprehensive look at the open access movement, please see (Willinsky 2005)

³⁹ <http://www.thelatinlibrary.com/>.

⁴⁰ <http://www.thelatinlibrary.com/readme>.

⁴¹ <http://www.perseus.tufts.edu/hopper/>

attribution/share-alike/non-commercial license.⁴² Third parties may thus freely create new resources based on what we provide but they must make their additions available under the same terms and they cannot restrict access to these resources behind a subscription barrier. The non-commercial license does not exclude advertising based revenue and we hope that internet services such as Google, Yahoo and Microsoft will load everything that we produce into their collections.

Since spring 2005, we have provided a Web service that exposes well-formed chunks of our data to third parties. In March 2006, we have made available under a Creative Commons license the TEI compliant XML files for the Greek and Latin source texts that we have created that were based upon public domain editions: c. 13,000,000 words of text. While this collection is much smaller than the 76,000,000 words on the 2000 TLG E Disk or the 91,000,000 words on the spring 2006 TLG Website, it does already contain most of classical Greek and many classical Latin source texts. All of our unencumbered lexica, encyclopedias, commentaries, and other reference materials will follow suit and be released under the same license. Likewise, all components of the new digital library system that underlies Perseus are being written for open source distribution and will, we hope, be integrated into the next generation of digital library systems.

Hypertextual Writing

As with access, hypertextual documents depend upon policy – even Web links, primitive though they may be, provide a starting point. The classicist Christopher Blackwell has produced what may be the best example of a publication that bridges the gap between traditional print and densely hypertextual Web publication. He produced an electronic publication as his tenure book, a web site that surveys Athenian democracy.⁴³ Figure 2 illustrates a snapshot of this site. The site includes not only PDF visualizations of the text optimized for print but also HTML representations of the same documents. The HTML documents contain a dense set of primary source citations that are filtered out of the print-oriented PDF publications. Blackwell has striven to provide the primary source evidence behind every significant assertion. The secondary scholarship on this subject has grown so tangled that many publications simply cite other secondary scholarship, leaving readers to dig through multiple sources before they can assess the underlying evidence. Blackwell's publication assumes the presence of a stable, comprehensive digital library to make the citations actionable links.

⁴² <http://creativecommons.org/>

⁴³ <http://www.stoa.org/projects/demos/home>

The screenshot displays a digital text interface. On the left is a vertical table of contents with links such as 'Demos Home', 'Summary', 'General Principles', 'Eligibility and Selection', 'Scrutiny of Councilors', 'The Bouleutic Oath', 'Presidents and Chairman', 'Rewards for Service', 'Times and Places of Meetings', 'Agenda for Meetings', 'Procedure for Meetings', 'Council Decrees', 'Independent Action', 'Introduction to Probouleumata', 'Exceptional Decrees', 'Probouleumata Voted Down', 'Open and Closed Probouleumata', 'Expiration of Probouleumata', 'Legislation', 'Jurisdiction', 'Powers to Punish', 'Administration of Attica', 'Public Finance', 'Foreign Policy', 'Secondary Works Cited', 'Index of Citations', and 'General Index'. The main text area is titled 'The Council' and is attributed to Christopher W. Blackwell, edition of January 23, 2003, page 2 of 24. The text discusses the Athenian democracy's three institutions: the courts (the People's Court and the Council of the Areopagus), the Assembly, and the Council (βουλή) (Dem. 20.100). It notes that at Athens, the Council was formally called the Council of the 500 (ἡ βουλή οἱ πεντακόσιοι), to distinguish it from the Council of the Areopagus (see, for example, Dem. 19.179; SEG 19 133). A pop-up window titled 'Summary: The Council of the Areopagus' is overlaid on the text, providing a summary of the Council of the Areopagus, its location (the Areopagus, or 'Hill of Ares'), and its role as a legal institution under the Athenian democracy.

Figure 2: Hypertextual writing from Christopher Blackwell's *Demos*, which illustrates a genuine step beyond print monographs.

Hypertextual writing builds on ubiquitous access to source materials. We can create hypertextual documents with links to subscription-based resources, but in so doing we implicitly define an audience of academics and a handful of committed non-professionals with access to good libraries. Hypertextual writing hidden from the outside world behind subscription barriers cannot, of course, reach beyond academic elites. Dense hypertextual links that are in open-access publications but that point to academic subscription based sources have no more impact on society as a whole than citations to print-only resources. Only open access publications with links to open access sources can increase the transparency of what we in the humanities do and engage a broader audience in the intellectual discourse that we pursue.⁴⁴

Aside from the content, Blackwell's work demonstrates the potential of the form and exhibits a scholarly leadership badly needed within the humanities. Had he worked with a conventional academic publisher he might have earned greater conventional prestige,

⁴⁴ The work of the Public Knowledge Project attempts to link scholarship to freely available sources in order to support reading by a broader audience; see (Willinsky 2003).

but he would have reached a smaller audience and would probably not have had the freedom to create expository texts so well adapted to the digital environment.

Fine grained, repurposable digital objects

We need compound documents, dynamically generated to serve particular users at particular times, that draw upon materials from a range of sources to create a new, unified whole.⁴⁵ Such documents have two requirements:

```
- <TEI.2>
  - <text>
    - <body>
      - <div0 n="*s111" type="alphabetic letter" org="uniform" sample="complete">
        - <entryFree id="n102117" key="sw/frwn" type="main" opt="n">
          <orth extent="full" lang="greek" opt="n">SW/FRWN</orth>
          ,
          - <gramGrp opt="n">
            <gram type="dialect" opt="n">Ep.</gram>
          </gramGrp>
          and poet.
          <orth extent="full" lang="greek" opt="n">SA^O/FRWN</orth>
          (as in
          - <bibl default="NO">
            <author>Hom.</author>
          </bibl>
          ,
          <abbr>v.</abbr>
          infr.,
          - <bibl n="Perseus:abo:tlg,0033,005:9:46" default="NO">
            <author>Pi.</author>
            <title>Pac.</title>
            <biblScope>9.46</biblScope>
          </bibl>
          ),
          <orth extent="full" lang="greek" opt="n">ONOS</orth>
          ,
          <gen lang="greek" opt="n">O</gen>
          <foreign lang="greek">, H</foreign>
          : neut.
          <foreign lang="greek">SW=FRON</foreign>
          :—prop.
          - <sense id="n102117.0" n="A" level="1" opt="n">
            <tr opt="n">of sound mind</tr>
```

Figure 3: XML Entry from LSJ 9 on the Perseus Website

- Rights agreements that provide access to source objects and their constituent parts (e.g., TEI XML, the measurements underlying a 3D model) rather than their derivatives (e.g., HTML, Quicktime VR). This reflects a simple, but profound, commitment that differs from the rights regimes that predominate in the Web.

⁴⁵ This need for reusable digital objects that can draw upon a range of services is a major theme of the recent Mellon funded study on support interoperability between digital repositories (Bekaert 2006).

- Well-structured source objects: Access to the digital text of a dictionary does us little good if the text does not mark the headwords and the beginnings and the senses and other components of individual articles.⁴⁶ Most SGML/XML documents available on-line have very simple structures that do not capture crucial data (e.g., the entries in the book index, which allow us to draw on human, rather than machine, decisions as to whether a particular Salamis is part of Athens or Cyprus).⁴⁷

Figure 3 illustrates an entry from the Liddell Scott Jones Greek English Lexicon (LSJ 9)⁴⁸ Notice that the mention of “Pi. Pae.” has not been expanded to a textual form but has been linked instead to an authority list (in this case, the numeration of the TLG Canon⁴⁹) unambiguously stating that “Pi. Pae.” denotes Pindar’s *Paeon odes*. Such links are fundamental as collections grow larger and increasingly ambiguous. The beginnings and ends, not only of the article as a whole but of each sense within it, are clearly marked and each has a unique identifier with which other documents can cite it.

Third parties can dynamically extract well-formed fragments of XML from the Perseus Digital Library, including canonical chunks of source texts, articles from various reference works, as well as the entire contents or individual senses from lexica. Figure 4 shows the same article as it appears in <http://www.dendrea.org/>, a third party site separate from the Perseus source collection: because it has access to the XML source, this site has been able to generate services (such as a browser for etymologically related terms, synonyms and antonyms) not available at Perseus.

ΔΕΝΔΡΕΑ

configure
contact

Etym	Syn	Ant	Rel
ΣΩΣ-			
συνδιασώζω			
σώζω			
ἀνασώζω			
ἀποσώζω			
διασώζω			
ἐκσώζω			
σώσ			
σωτήρ			
σωτηρία			
σωτήριος			
σωφρονέω			
σωφρονίζω			
σωφρονικός			
σωφροσύνη			
σωφρόσυτος			
σώφρων			

σώφρων, Ep. and poet. σα⁴όφρων (as in *Hom.*, v. infr., *Pi.Pae.9.46*), ονος, ὁ, ἡ; neut. *σώφρων*:—prop.

A. of sound mind (from *σῶς*, φρήν, cf. *Pl.Cra.411e*, *Arist.EN1140b11*): hence, **discreet, prudent**, "οὐκ ἄν με σαόφρονα μυθήσαιο ἔμμεναι" *Il.21.462*, cf. *Od.4.158*; opp. ἄφρων, *Thgn.431, 454, 497*; opp. νήπιος, *Id.483*; opp. ἀνόητος, *Hdt.1.4*; "σώφρονες περὶ θεός" *X.Mem.4.3.2*; "σωφρονέστατος ἐν τῇ τέχνῃ" *Hp.Promth. 2.2*.

2. of things, "τοῖσι λόγοις σώφρον ἔπεισιν ἄνθος" *Ar.Nu.1025* (lyr.); ζ. οἰκτος **reasonable** compassion, *Th.3.59*; "-έστατον κήρυγμα" *Aeschin.3.4*; "σώφρον' εἶπας" *E.IA1024*; "ἄλλο τι -έστερον γνώσεσθε" *Th.5.111*; *σώφρον ἐστι* c. inf., *Id.1.42*.

II. in Att., esp. having control over the sensual desires, temperate, self-controlled, chaste ("σώφρων ὁ μετρίας ἐπιθυμίας ἔχων" *Pl.Def.415d*, cf. "σωφροσύνη" 1), "μοι δὸς -εστέραν πολὺν μητρὸς γενέσθαι" *A.Ch.140*, cf. *S.Aj.132*; γυνή ζ. *And. 4.14*, cf. *S.Fr.682*; "ζ. καὶ ἐγκρατῆς ἑαυτοῦ" *Pl.Grg.491d*, cf. 1 *Ep.Ti. 3.2*, etc.

⁴⁶ For a good overview of the possibilities inherent in better exploiting the semantic content of digital objects, please see (Bearman and Trant, 2005).

⁴⁷ A similar issue is often raised by those researchers who wish to analyze Wikipedia, but find its unstructured data requires a great deal of work to support automated processing. See (Volkel 2006).

⁴⁸ (Liddell et al. 1940).

⁴⁹ (Berkowitz and Squitier 1990.)

Figure 4: LSJ Entry from Dendrea website

Documents that learn from each other

The artificial intelligence pioneer Marvin Minsky suggested that the time would come when no one will imagine that the books in a library did not talk with one another. While Minsky may have envisioned very powerful artificial intelligence spawning conversations between books far beyond what is currently possible, our books are already beginning to converse in simple but substantive ways.⁵⁰ Put another way, so much material is already on-line that only machines can scan more than a tiny fraction of what is available. Smart books are already beginning to appear to provide knowledge-intensive services and offer up more information about themselves than any reader might have thought to ask.

Figures 5, 6, 7 and 8 illustrate four dynamically generated views based on the interaction of different books within the Perseus digital collection.

Thucydides, *The Peloponnesian War*

Your current position in the text is marked in blue. Click anywhere in the line to jump to another position: [Hide browse bar](#)

book: chapter:

This text is part of: [Greek and Roman Materials](#)

View text chunked by: [book](#) : [chapter](#) : [section](#)

Table of Contents:

- ▼ [Book 1](#)
- [chapter 1](#)
- [chapter 2](#)
- [chapter 3](#)
- [chapter 4](#)
- [chapter 5](#)
- [chapter 6](#)
- [chapter 7](#)
- [chapter 8](#)
- [chapter 9](#)
- [chapter 10](#)
- [chapter 11](#)

86.

"The long speech of the Athenians I do not pretend to understand. They said a good deal in praise of themselves, but nowhere denied that they are injuring our allies and Peloponnesians. And yet if they behaved well against the Medes then, but ill towards us now, they deserve double punishment for having ceased to be good and for having become bad. [2] We meanwhile are the same then and now, and shall not, if we are wise, disregard the wrongs of our allies, or put off till tomorrow the duty of assisting those who must suffer to-day. [3] Others have much money and ships and horses, but we have good allies whom we must not give up to the Athenians, nor by lawsuits and words decide the matter, as it is anything but in word that we are harmed, but render instant and powerful help. [4] And let us not be told that it is fitting for us to deliberate under injustice; long deliberation is rather fitting for those who have injustice in contemplation. [5] Vote therefore, Lacedaemonians, for war, as the honor of Sparta demands, and neither allow the further

Greek [focus](#) [load](#)

English (Thomas Hobbes) [focus](#) [load](#)

English (Benjamin Jowett) [focus](#) [load](#)

Notes (E. C. Marchant) [focus](#) [load](#)

Notes (Charles D. Morris) [focus](#) [load](#)

Places (automatically extracted) [hide](#)

Sort places [alphabetically](#), [as they appear on the page](#), [by frequency](#)
Click on a place to search for it in this document. [More about these results...](#)

- [Sparta \(Canada\)](#) (1)
- [Peloponnesians \(Greece\)](#) (1)
- [Mede \(Italy\)](#) (1)
- [Athens \(Alabama, United States\)](#) (1)

References [hide](#)

Found 31 references related to this page.

- Cross-references to this page (16):
 - Herbert Weir Smyth, *A Greek Grammar for Colleges*, [DATIVE OF INTEREST](#)
 - Herbert Weir Smyth, *A Greek Grammar for Colleges*, [NEGATIVE \(PROHIBITIONS\)](#)
 - Herbert Weir Smyth, *A Greek Grammar for Colleges*, [VERBAL ADJECTIVES IN -ιστος](#)
 - Raphael Kühner, Bernhard Gerth, *Ausführliche*

Figure 5: Basic Report: A user has called up a translation of Thucydides, *History of the Peloponnesian War*, Book 1, chapter 86.

Figure 5 is a “basic report” from the Perseus website that lists various translations, editions, commentaries and other resources about a particular passage of classical Greek

⁵⁰ For more on the potential of what can happen when the knowledge within digitized books interacts, please see (Kelly 2006), (Crane 2005a), (Crane 2005b).

— Thucydides' *History of the Peloponnesian War*, Book 1, chapter 86. While it resembles the page of a book, it reflects the fact that many books have been analyzed and relevant sections extracted to create a dynamic view that would be not feasible in print. Different works represent Thucydides as “Th.”, “Thuc.”, “T.”, “Thucyd.”, etc., the history as “Hist.”, “H.”, “Pel. War”, etc., and the citation as “I, 86”, “I.86”, “1,86”, “1.86”. All of these representations are mapped onto a single canonical reference around which we can then cluster a range of information. When the user calls up one translation, the translation calls out to the library for other translations, Greek editions, commentaries, lexica, grammars and other reference works which cite words in this passage. The text in focus thus interacts with a range of other related resources, which align themselves in real time, ready to provide background information or to become themselves the focus of attention.

Figure 6 displays the word clusters associated with uses of the Greek word *arche* in Thucydides' *History of the Peloponnesian War* (c. 150,000 words) and five English translations. By comparing the English translations with the source text, the automatic process identified clusters of meaning associated with various Greek words – in effect, creating a rough English/Greek lexicon and semantic network. The clusters capture the senses “empire,” “government,” “political office,” and “beginning.” The cluster headed “ancient” (marked in bold) captures a distinct word that happens to share the stem *arch*. Such parallel text analysis can update its results as new translations and source texts appear within the system, providing dynamic conclusions based on interaction of books within the digital library.

<u>empire</u>	<u>dominion</u>	<u>power</u>	<u>government</u>
<u>office</u>	<u>government</u>	<u>magistrates</u>	people
<u>command</u>	Mindarus	Tissaphernes	Laches
<u>power</u>	Eurystheus	king	Atreus
<u>dominion</u>	<u>power</u>	<u>rule</u>	Hellenes
<u>magistrates</u>	Theseus	people	council
<u>government</u>	<u>power</u>	Hippias	Pharnabazus
ancient	descendants	temples	Pythian
whom	<u>beginning</u>	pits	just
called	Zancle	Pangaeus	<u>originally</u>
Harmodius	<u>originally</u>	basket	Cyclopes
Philip	brother	<u>government</u>	Sitalces

Figure 6: Parallel text analysis: Word clusters associated with uses of the Greek word *arche* in Thucydides' *History of the Peloponnesian War* (c. 150,000 words) and five English translations. Translation equivalents are underlined.⁵¹

Likewise, Figure 7 shows the results of automatic named entity identification. In this case, a translation of Thucydides compares its vocabulary to authority lists such as

⁵¹ This work was done by D. Sculley, Phd candidate in Computer Science at Tufts University.

encyclopedias and gazetteers to determine possible names and then uses the context in other books to resolve ambiguous references in actual text⁵² (e.g., does “Salamis” designate the island near Athens, a place in Cyprus or some other location)?

```

9 <milestone unit="sentence" n="974"/></seg> </p>
0 <p> <milestone n="108" unit="chapter"/> <milestone unit="section" n="1"/>
0 <seg> <milestone unit="para" ed="P"/>About the same time <persName
0 n="Alcibiades,,,,," id="n-0001.0000.00000.01912"
0 reg="mostcommon:Alcibiades,nomatch:0"><surname>Alcibiades</surname></persName
0 > returned with his <num value="13">thirteen</num> ships from <placeName
0 key="perseus,Caunus">Caunus</placeName> and <placeName reg="Phaselis, Antalya
0 Ili, Akdeniz kiyisi" key="tgn,7002612">Phaselis</placeName> to <placeName
0 key="tgn,7002673">Samos</placeName>, bringing word that he had prevented the

```

Figure 7: Named Entity Tagging: An XML fragment of Thucydides with all named entities automatically extracted and disambiguated.

Figure 8 shows the results of automatic syntactic parsing. Here a parser assigns tags to words by comparing the current text to other texts that have been syntactically analyzed by hand. By communicating with other texts in this way, the parser can determine the likelihood that a given morphological sequence (e.g., accusative noun, accusative noun, preposition, ablative noun) has a given syntactic parse. In the prototype shown in the two figures below, only tags with a reasonably high probability are assigned (allowing the system to have higher precision at the expense of greater coverage). If errors arise (as below, where *Romam* should not modify *Vrbem* as an apposition), users can correct the syntactic dependencies to improve the overall system, providing a valuable feedback mechanism whereby both the user and the text can productively learn from each other.

⁵² (Smith 2001). For more on the technical details of this system, see (Crane and Jones 2005).

Cornelius Tacitus, *Annales*

Agamemnon Search ("Agamemnon", "Hom. Od. 9.1", "denarius") [advanced search] [view abbreviations]

Your current position in the text is marked in blue. Click anywhere in the line to jump to another position: [Hide browse bar](#)

book: _____

This text is part of: [Table of Contents](#) Tac. Ann. 1.1

Click on a word to bring up parses, dictionary entries, and frequency statistics

1. Vrbem Romam a principio reges habuere; libertatem et consulatum L. Brutus instituit. dictaturae ad tempus sumebantur; neque decemviralis potestas ultra biennium, neque tribunorum militum consulare ius diu valuit. non Cinnae, non Sullae longa dominatio; et Pompei Crassique potentia cito in Caesarem, Lepidi atque Antonii arma in Augustum cessere, qui cuncta discordiis civilibus fessa nomine principis sub imperium accepit. sed veteris populi Romani prospera vel adversa claris scriptoribus memorata

References (17 total) [show](#)

Vocabulary Tool [load](#)

Syntax [hide](#)

See a syntactic parse of this sentence:

- Vrbem Romam a principio reges habuere
- libertatem et consulatum L. Brutus instituit
- dictaturae ad tempus sumebantur
- neque decemviralis potestas ultra biennium, neque tribunorum militum consulare ius diu valuit
- non Cinnae, non Sullae longa dominatio
- et Pompei Crassique potentia cito in Caesarem, Lepidi atque Antonii arma in Augustum cessere, qui cuncta discordiis civilibus fessa nomine principis sub imperium accepit
- sed veteris populi Romani prospera vel adversa claris scriptoribus memorata sunt
- temporibusque Augusti dicendis non defuere decora ingenia, donec gliscente adulatione detererentur
- Tiberii Gaique et Claudii ac Neronis res florentibus ipsis ob metum falsae, postquam occiderant recentibus odiis compositae sunt
- inde consilium mihi pauca de Augusto et extrema tradere, mox Tiberii principatum et cetera, sine ira et studio, quorum causas procul habeo

Latin Dependency Treebank

Vrbem Romam a principio reges habuere

index	word	head	relation	lemma + morph	add new lemma	add new morph
0	Vrbem			noun sg fem acc		
1	Romam	0	APOS	noun sg fem acc		
2	a	5	ADV	prep		
3	principio	2	J	noun sg neut abl		
4	reges	5	SBJ	noun pl masc nom		
5	habuere			verb 3rd pl perf ind act		

Update Save

Vrbem Romam a principio reges habuere
 +-----ADV-----+
 +<APOS+
 +-<J--+
 +SBJ>--+
 Vrbem Romam a principio reges habuere

Search [hide](#)

Search in Latin. [More search options](#)

Search to:

- All Collections
- Greek and Roman Materials
- Latin Prose
- Latin Texts
- Tacitus
- Tacitus, Annales
- Annales* (this document)

Search for all inflected forms
 Search for "amo" returns "amo", "amas", "amat", etc.)
 Search for exact forms only

Display Preferences [hide](#)

Text Display: Unicode (precombined)

Figure 8: A prototype of a basic report of Tacitus' *Annales* where users have the option to see automatically generated syntactic parses of the sentences. Users can contribute to the system by correcting the automatic parse (e.g., *Romam* should not be in apposition to *Vrbem*) and transforming the partial parse into a complete one (here, by assigning tags to *Vrbem* and *habuere*).

The figures above thus provide initial examples of books interacting with each other to create new forms of publication. These examples point the way towards increasingly intelligent collections which become more powerful and sophisticated as their size and internal structure improve – the more books communicate with each other, the more information about themselves they can provide.

Documents that learn from their audiences

Documents can learn from each other and drive automated processes to identify people and places in full text, analyze the contents of collections to provide integrated reports drawing on multiple information sources and perform similar tasks to apply existing classification or mine new potential knowledge.⁵³ But even when such processes address

⁵³ A variety of work is beginning to explore how to best exploit both the structured and unstructured knowledge already present in digital library collections to train other systems with document analysis and machine learning; see for example (Nagy and Lopresti 2005) and (Esposito, et. al. 2005).

questions with discrete, decidable answers, users will want to refine the results and these user-contributed refinements are important not only for other users but for improving the quality of subsequent automated analysis.⁵⁴ Thus, an automated system may incorrectly identify “Washington” in one passage as Washington, DC, when it is in fact Washington state. Or it may simply fail because its gazetteer does not include an entry for the right Washington in a given passage (e.g., Washington, NC). Thus, even when working with very simple conceptual systems, users should be able to correct system conclusions whether by selecting a different existing answer or by adding a new possible answer to the existing set. Figure 9 shows an existing feedback mechanism whereby users can vote against a machine generated analysis.

As machines perform more sophisticated analyses where there is no single right answer, user feedback may be even more important: lexicographers do not always agree on how to describe the senses of a word.⁵⁵ Machines can infer possible senses by studying the contexts in which a word appears but we still want to be able to modify the suggested word senses, even if experienced lexicographers would not agree on any one final configuration of senses.

P. Vergilius Maro, *Aeneid*
J. B. Greenough, Ed.

Your current position in the text is marked in blue. Click anywhere in the line to jump to that position.

book: _____
line: _____

This text is part of:
[Greek and Roman Materials](#)
[Latin Poetry](#)
[Latin Texts](#)
[Vergil](#)
[Vergil, *Aeneid*](#)

View text chunked by:
[book : line](#)
[book : line](#)

Table of Contents:
[Book 1](#)
[Book 2](#)
[Book 3](#)
[Book 4](#)
[line 1](#)
[line 31](#)
[line 54](#)
[line 90](#)

Click on a word to bring up parses, dictionary entries and statistics

At regina gravi iamdudum saucia cura
volnus alit venis, et caeco carpitur igni
Multa viri virtus animo, multusque re-
gentis honos: haerent infixi pectore vo-
verbaque, nec placidam membris dat c-
Postera Phoebæa lustrabat lampade te-
umentemque Aurora polo dimoverat t-
cum sic unanimam adloquitur male sa-
“Anna soror, quæ me suspensam inso-
Quis novus hic nostris successit sedibi-
quem sese ore ferens, quam forti pect-
Credo equidem, nec vana fides, genus
Degeneres animos timor arguit: heu, c-
iactatus fati! Quæ bella exhausta can-
Si mihi non animo fixum immotumque
ne cui me vinco vellem sociare iugali,
postquam primus amor deceptam moi-
si non pertaesum thalami taedæque fi-
huic uni forsân potui succumbere culp-
Anna, fatebor enim, miseri post fata S-
coniugis et sparsos fraterna caede Pen-

Word Study Tool
Get Info for _____ in Latin Go

saucius
(Show lexicon entry in [Elem. Lewis Lewis & Short](#)) (search)

saucia	adj pl neut nom	no user votes	14.1%	[vote]
saucia	adj pl neut voc	no user votes	13.8%	[vote]
saucia	adj pl neut acc	no user votes	13.9%	[vote]
saucia_	adj sg fem abl	no user votes	13.9%	[vote]
saucia	adj sg fem nom	no user votes	14%	[vote]
saucia	adj sg fem voc	no user votes	13.6%	[vote]

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
609375	33	0.54	18	0.30	Latin Poetry
3414041	99	0.29	71	0.21	Latin Texts
83620	8	0.96	7	0.84	Vergil
63770	8	1.25	7	1.10	P. Vergilius Maro, <i>Aeneid</i>

saucio to wound, hurt
(Show lexicon entry in [Elem. Lewis Lewis & Short](#)) (search)

saucia_ † verb 2nd sg pres imperat act no user votes 16.6% [vote]

† This form has been selected using statistical methods as the most likely one in this context. It may or may not be the correct form. (More info)

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
609375	19	0.31	4	0.07	Latin Poetry
3414041	42	0.12	14	0.04	Latin Texts

⁵⁴ Research into how to capture the knowledge of users to drive both machine learning processes and personalization is growing rapidly, see for example (Chklovski 2005), (Carrera 2005) (Gilardoni 2005), (Kruk 2005).

⁵⁵ Some initial work in having user contributions assist in automated word sense disambiguation has been reported in (Navigli and Velardi 2005).

Figure 9: A morphological analysis system: This system has calculated the possible analyses for a given form. A simple machine learning system has ranked the possibilities of each analysis in the given context. Users can now vote for the analysis which they see as correct.

Documents that adapt themselves to their users

Customization and personalization constitute two other methods by which machines respond dynamically to user behavior. In customization, users explicitly set parameters to shape subsequent system behavior. Personalization generally implies that the system takes action on its own, comparing the behavior of a new user to that of other users that it has encountered in the past.⁵⁶ Some of us create our own customized versions of internet portals (e.g., “My Yahoo!”). Most humanists have, by 2006, encountered the personalization on sites such as Amazon, which inform us that people who bought the book that we just chose also bought books X, Y, and Z.⁵⁷

Both customization and personalization have great potential within the humanities.⁵⁸

C. Suetonius Tranquillus, *Caligula*
Maximilian Ihm, Ed.

[Study vocabulary in this passage.](#)

Table of Contents [←](#) [→](#)

Click on a word to bring up parses, dictionary entries

This text is part of:
[Greek and Roman Materials](#)
[Latin Prose](#)
[Latin Texts](#)
[Suetonius](#)

View text chunked by:
[life](#) : [chapter](#) : [section](#)

Table of Contents:
[▶ Divus Iulius](#)
[▶ Divus Augustus](#)

[XML](#) [←](#) [→](#)

Your vocabulary profile:
Wheelock (5th)
Wheelock, Frederick M., [Wheelock's Latin \(5th Edition\)](#) (1990)

This passage contains **115** possible dictionary forms.
According to your vocabulary profile, you have already learned **54** of
This page displays the **61** remaining dictionary forms.

[Customize your vocabulary profile](#)

Frequency	Dictionary Form	Short Definition
2	contio	a meeting, assembly, convocation, gathering, audience
1	acerbitas	bitterness, harshness, sourness
1	armatus	armed, equipped, in arms
1	armo	to furnish with weapons, arm, equip
1	atrocitas	fierceness, harshness, enormity
1	augustus	consecrated, sacred, reverend
1	Augustus	a cognomen given to Octavius Caesar as emperor, his majesty
1	circumdo	to place around, cause to surround, set around
1	cogitatio	a thinking, considering, deliberating, thought, reflection, meditation
1	confestim	immediately, speedily, without delay, forthwith, suddenly
1	contrucido	to cut to pieces, cut down, put to the sword
1	de	down (adv.)
1	decedo	to go away, depart, withdraw, retire

Figure 10: Customization in the Perseus Digital Library.⁵⁹

Figure 10 illustrates how a user profile can help filter information, showing readers what terms they have and have not encountered. A reader has informed the system that she has studied Latin from Wheelock’s fifth edition. The system has then compared a passage from Suetonius against the vocabulary in the textbook (drawing upon the morphological

⁵⁶ For an expansion of these definitions see (Russell 2003), and for a particular application (Bowen and Fantoni 2004).

⁵⁷ For more on the Amazon system, please see (Linden, et. al. 2003)

⁵⁸ There is growing body of literature as to how these technologies might be applied within the humanities, most often digital libraries, for an overview please see (Smeaton and Callan 2005),

⁵⁹ This work was done by David Mimno and Gabriel Weaver, Perseus Project, Tufts University.

analysis system which can match inflected words to their dictionary entries). Of the 115 possible dictionary words in this passage, the reader has probably encountered 54 and will find 61 that are new. These new words are then listed according to their frequency in the given passage. Alternate sorting orders could stress words that would be important in readings that have been assigned for the rest of the semester, for Suetonius in general or for some particular topic (e.g., military events) of interest to the reader. The technology can be based on straightforward principles of ranking and filtering from information retrieval but have a significant impact. The example given addresses language learning but the same techniques are applicable to technical terms. The key to this approach would be the development of learning profiles which track the contents of many textbooks, handouts, and assigned readings over different learning which we pursue throughout our lives.⁶⁰

Figure 11 illustrates an example of personalization from the Perseus Digital Library. Once a user has asked for information on four or five words in a three hundred word passage of Ovid, we can then predict two thirds of the subsequent words that will elicit queries. This recommender system is similar in principle to the systems which Amazon and other e-commerce systems use to show consumers new products based on the products purchased by people who also bought product X. The application, however, reduces the search space of a language passage, suggesting words for study rather than products for purchase.⁶¹

P. Ovidius Naso, *Metamorphoses*
Hugo Magnus, Ed.

[Study vocabulary in this passage.](#) [Table of Contents](#) ← →

Click on a word to bring up parses, dictionary entries, statistics

Daphne.

Primus amor Phoebi Daphne Peneia, quem fors ignara dedit, sed saeva Cupidinis ira. Delius hunc, nuper victa serpente superbus, viderat adducto flectentem cornua nervo “quid” que “tibi, lascive puer, cum fortibus dixerat, “ista decent umeros gestamina nos qui dare certa ferae, dare vulnera possumus qui modo pestifero tot iugera ventre preme stravimus innumeris tumidum Pythona sagi Tu face nescio quos esto contentus amores inritare tua, nec laudes adserere nostras.”

Table of Contents: Filius huic Veneris “figat tuos omnia, Phoeb

Word Study Tool

Get Info for in Latin

dico to say, speak, utter, tell, mention, relate, affirm, decide, assert

(Show lexicon entry in [Elem. Lewis Lewis & Short](#)) ([search](#))

dixerat verb 3rd sg plup ind act

Based on the words in this passage for which you have received information, you may be interested in the following words:

[ista](#) “this, that, he, she”
[umeros](#) “the upper arm, shoulder”
[ferae](#) “a wild beast, wild animal” or “wild, untamed, uncultivated”
[vulnera](#) “a wound” or “to wound”
[prementem](#) “to press”

Word Frequency Statistics ([more statistics](#))

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
609375	1710	28.06	915	15.02	Latin Poetry
3414041	22613	66.24	16610	48.65	Latin Texts
140315	587	41.83	395	28.15	Ovid

[XML](#)

⁶⁰ Developing accurate user models and profiles to support and track learning is a topic of significant study, for some recent work please see (Brusilovsky 2005) and (Kavcic 2004).

⁶¹ Work on how personalization, particularly recommender systems, might be used within humanities environments has been explored by (Bia 2004), (Kim 2004), to name only a few.

Figure 11: Personalization in the Perseus Digital Library.⁶²

Customization and personalization are fundamental technologies. While the examples given above address the needs of intermediate language learning, the same techniques would support professional researchers working with source materials outside of their own areas of specialization (e.g., an English professor with a background in classical Latin working through 16th century English Latin prose). Customization and personalization have potential for filtering and structuring information for experts within their own field of expertise. They are core services for any advanced digital infrastructure underlying ePhilology.

Building the infrastructure for ePhilology

The examples in the preceding section illustrate current steps towards future possibilities. This section describes an infrastructure to move the field forward. On the one hand, we need to exploit emerging technologies. This not only includes downloading applications and compiling source code but reading research publications and implementing suitable algorithms. At the same time, in the long run we in classics and in the humanities may primarily contribute the knowledge sources whereby developed tools can analyze historical materials. Thus, named entity systems applied to texts about the Greco-Roman world will perform much better if they have access to information about the people and places of the Greco-Roman world than if they must rely wholly on resources which describe the contemporary world.⁶³

Primary sources and reference materials that evolve in real time should include the following features:

- **Open source/polyphonic:** we need encoded knowledge that can be maintained in real time and that can incorporate multiple points of view. Core resources should not be held restricted by rights agreements but should serve as a common resource to which others may add and from which others may generate new resources.⁶⁴ New variations on traditional review will emerge, with on-going usage within the scholarly community complementing – and perhaps in some measure supplanting – the hit or miss preparatory edits of static documents necessary for print. In a digital world, capital information sources such as editions and reference works evolve: where print publication freezes documents, digital publication only begins its functional life after publication. We can easily preserve versions of the document as it appeared at any one time (thus allowing us to see what an author saw when the citation was added to an argument) and track who contributed what and at what time.

⁶² This work was done by D. Sculley, Phd candidate in Computer Science at Tufts University under the supervision of Professor Carla Brodley, with help from Gabriel Weaver of the Perseus Project.

⁶³ On the need for historical knowledge sources, see (Crane and Jones 2006b) and also (Siemens 2006)

⁶⁴ For some recent work on creating reference works that allow users to both edit and create materials, please see (Witte 2005) and (Kolbistch 2005).

- Readable by machines and people alike: Our dictionaries should be able to search new texts for the varying senses claimed for each word; our encyclopedias should scan secondary sources for, and then summarize the results of, new discussions of the people, places and topics which they cover; our texts should collate themselves against other witnesses and editions as these come on-line. The more machines can understand, the more effectively they will be able to support the questions that we pose and to provide the personalized background that we need.⁶⁵ The need to add the greater structure and consistency needed for machine processing only highlights the need for materials that we can freely reformat.

These features have at least one profound implication. Once documents become dynamic and can evolve over time, we must evaluate them according to their potential for growth – their state at any one time constitutes only a single data point. In classics, editions and reference works more than a century old but which are in the public domain and can be freely updated may thus prove more valuable in an electronic environment than the best current resources if these are either static or even updated according to a traditional editorial process.

A range of community driven reference works has emerged in recent years. The most famous, Wikipedia, arguably constitutes the most important intellectual development of the early twenty-first century: a new form of intellectual production, community driven and dynamic, has produced more than 1,000,000 general articles in five years.⁶⁶ If and when the need for new articles diminishes, it will be interesting to see whether this vast resource enters a phase of refinement, thus suggesting a two-fold model: an open phase of development to bootstrap the system, followed by a period of revision. Criticizing this work is important, but only insofar as such criticism helps us to draw upon and contribute to this flood of intellectual energy.⁶⁷ Other community driven systems with more centralized editorial control have appeared for math and physics.⁶⁸ A 2005 grant from the National Endowment for the Humanities has even provided support for Pleiades, a community driven project on Greco-Roman geography.⁶⁹

An infrastructure for ePhilology would contain two fundamental components: the primary sources and a network of reference works, linked to and constructed from the sources. Dynamic and intelligent links should connect all components of the infrastructure. When changes are suggested to a text, the effects of these changes upon associated reference works should be tracked and all affected places in all reference works should automatically report the change. Conversely, work based on analysis of a particular reference work should be noted in the text (e.g., a new study of a particular person that suggests reading one name vs. another).

⁶⁵ For an intriguing exploration of the potential of “machines as readers”, see (Shamos 2005).

⁶⁶ As of May 23, 2006, the count for English articles on <http://www.wikipedia.org> stands at 1,145,000.

⁶⁷ For example, see (Rosenzweig 2006)

⁶⁸ <http://planetmath.org/>; <http://planetphysics.org/>.

⁶⁹ <http://www.unc.edu/awmc/pleiades.html>.

Technically, this environment needs two things: a set of data structures and data. The Text Encoding Initiative provides serviceable structures for texts themselves.⁷⁰ Text mining can identify many patterns latent within these texts,⁷¹ but once we have ways of identifying people, places, organizations and other entities within texts we need methods to reason, at least in rudimentary fashion, about them. Knowledge bases differ from databases in that they are designed to support inferencing: thus, if the system knows that no events in Herodotus postdate 400 BCE, that Alexander the Great was born after 400 BCE and that Alexander the Great was a king of Macedon, then it can avoid identifying the Alexander, king of Macedon, in Herodotus as Alexander the Great. Fortunately, the slowly emerging Semantic Web is designed to support such reasoning. Promising formats exist for geographic information⁷² and for museum objects,⁷³ and we now have a well-developed set of guidelines for ontology production in OWL (Web Ontology Language).⁷⁴ Ontologies, however, rapidly grow idiosyncratic and their development is as much a social as a technical process.⁷⁵ To drive that development, however, we need enough data for serious experimentation – data structures and data will need to evolve, however cautiously, in tandem. We need services of interest to attract long term user communities and enough data to raise issues of scale if we are to engineer solutions that will support intellectual life over time.

The Google Library and especially the Open Content Alliance, which has an open source policy, will help provide access to image books of virtually all useful public domain materials. These will provide immediate access to Latin and Roman script publications, with searchable OCR for classical Greek probably not far behind. These texts will provide the foundation on which we can build a dynamic knowledge base that evolves and grows more intelligent.

Moving from print to knowledge involves three steps:

- 1) Initial markup to capture the basic structural elements: we need the headwords for dictionaries/lexica/gazetteers, clear separation of headers, footnotes and text, and other basic elements not present in raw OCR.⁷⁶
- 2) Semantic analysis: classification of proper names (e.g., is Peneius the river or the river god?) and identification of basic propositional statements (e.g., “a REGION of PLACE,” “PERSON born at PLACE in DATE”).⁷⁷

⁷⁰ <http://www.tei-c.org/>

⁷¹ This is the approach of the Nora text mining project: <http://nora.lis.uiuc.edu/description.php>; (Plaisant et al. 2006).

⁷² <http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2-guide.htm>.

⁷³ <http://cidoc.ics.forth.gr/>.

⁷⁴ <http://www.w3.org/TR/owl-features/>.

⁷⁵ For some particular applications of ontologies in the humanities, see (Nagypal 2004), (Nagypal 2005), (Mirzaee 2005), and for the merging of various efforts, see (Eide 2006) and (Doerr 2003).

⁷⁶ For some lengthier discussion of these issues see (Bearman and Trant 2005) and (Sankar 2006).

⁷⁷ Named entity recognition and semantic classification have large bodies of literature, but the use of these applications in the humanities is receiving more examination see (Hoekstra 2005) and (Shoemaker 2005)

- 3) Alignment against pre-existing entries common list and identification of new entries: Alexander-12 in encyclopedia-1 may be equivalent to Alexander-32 in encyclopedia-2 or it may represent an entirely new Alexander not yet attested.⁷⁸

Automated methods can address all three of the above phases but all methods are imperfect and print sources differ just enough that methods still need to be tuned for most reference materials. The three steps above constitute the most important and probably the most difficult work that we face, but they are essential and foundational to any serious infrastructure.

Classicists are fortunate in having a well-developed set of public domain print resources with which to begin their work.

- Texts: These can take older editions as their initial base texts but should then be (1) collated with other editions, both older and new, and (2) provide an initial database of variants and conjectures that can be expanded over time.⁷⁹ One well-tagged edition could help automatically identify and provide preliminary tagging for other on-line editions.⁸⁰ Perseus contains c. 70% of the corpus of classical and Hellenistic Greek and 50% of the corpus of classical Latin in TEI compliant XML. Both collections are expanding, with coverage of Latin being particularly cost effective: we should be able to provide coverage of 96% of the text on the PHI CD ROM, with later authors not in that collection (e.g., Ammianus, Sidonius) and a substantial postclassical collection.
- Translations: Scholars working with any historical language should, as a matter of principle, ensure that translations (1) are readily available and (2) flag those places where the new edition would impact at least one standard translation. Translations are, however, not only useful for those with little or no knowledge of the source language: parallel text analysis is a major component for machine translation (necessary where translations do not exist), automated lexicography, cross language information retrieval etc.⁸¹ Multiple translations of a single text strengthen statistical analysis. Translations are thus a high priority to any infrastructure for an ePhilology. In Perseus we have collected at least one translation for most of our sources.
- Morphology: The ability to connect a dictionary entry with its inflected forms is a fundamental service for any language. While the code needed to recognize legal combinations of stem and ending is challenging in Greek (where we must also consider augments, preverbs, accent and dialectics), morphological analysis is a

⁷⁸ For interesting work in this area, see (Barzilay 2005).

⁷⁹ For some previous work in this vein see (Spencer 2004).

⁸⁰ If we have “arma virumque cano Troiae qui primus ab oris” tagged in one text as Aen. 1.1, then we locate other instances of this line and apply the same markup. This strategy draws upon the fact that runs of repeated words are surprisingly uncommon, even in large corpora.

⁸¹ For a recent exploration of the uses of parallel texts, see (Mihalcea 2005), and their use in machine translation (Smith 2006).

data intensive process that depends upon lists of endings and especially stems. Since stems are, in practice, an unbounded set, assembling suitable databases of morphological data is the greatest challenge to morphological analysis in Latin and Greek. Dictionaries have provided the best general source for the stems, with Liddell, Scott, Jones (LSJ) and Lewis and Short helping us create databases with 52,700 Greek and 19,800 Latin stems. In 1990, we provided 100% coverage for the 1 million words of Greek included in Perseus 1.0. Many low frequency words and most proper nouns are not in these source lexica and only modest progress was made in extending this coverage. The need to improve morphological analysis provided one, though by no means the only, reason to identify, digitize and mine more comprehensive reference works with people and places.

- People and Places: For classical texts, the nineteenth-century three volume *Dictionary of Greek and Roman Biography and Mythology* (Smith 1873) and Smith's two-volume *Dictionary of Greek and Roman Geography* (Smith 1854) are more than a century older than the third edition of the *Oxford Classical Dictionary* (OCD3) (Hornblower and Spawforth 1996). Anyone looking for a survey of standard views from the late twentieth century must, of course, consult OCD3. Nevertheless, the older Smith dictionaries are better sources for ePhilology because they are more extensive and contain tens of thousands of machine extractable source citations. Both dictionaries set out, with reasonable success, to document all significant people and places mentioned in the literary corpus,⁸² with 20,000 and 10,000 entries in the biographical and geographical dictionaries. Equally important, we have been able to extract 37,500 and 25,800 citations, respectively. Each citation not only associates a particular passage with a particular topic but provides more materials whereby text-mining software can learn to distinguish the various Alexanders and Alexandrias when they appear elsewhere in primary and secondary sources alike. The original Smith articles can be mined for information about birth/death dates, family relations, place locations and other quantifiable data that can be used for intelligent information retrieval and general text mining.
- Authors, works, and their citation schemes: Authors comprise a key subset of people, with their works often listed in biographical entries of Smith's

⁸²(Smith 1873), p. ix: "Some difficulty has been experienced respecting the admission or rejection of certain names, but the following is the general principle which has been adopted. The names of all persons are inserted, who are mentioned in more than one passage of an ancient writer: but where a name occurs in only a single passage, and nothing more is known of the person than that passage contains, that name is in general omitted. On the other hand, the names of such persons are inserted when they are intimately connected with some great historical event, or there are other persons of the same name with whom they might be confounded"; (Smith 1854), p. viii: "Separate articles are given to the geographical names which occur in the chief classical authors, as well as to those which are found in the Geographers and Itineraries, wherever the latter are of importance in consequence of their connection with more celebrated names, or of their representing modern towns,—or from other causes. But it has been considered worse than useless to load the work with a barren list of names, many of them corrupt, and of which absolutely nothing is known. The reader, however, is not to conclude that a name is altogether omitted till he has consulted the Index; since in some cases an account is given, under other articles, of names which did not deserve a separate notice."

biographical dictionary, as well as good coverage for more than three and a half centuries of printed editions. The TLG and PHI Institute each have produced up-to-date catalogues of recent editions, as well as lists of author works. Lexica such as LSJ and Lewis and Short include extensive bibliographies of authors, works and older editions. While the Oxford Latin Dictionary is a relatively recent publication, it began work in the 1930s and the editions which it cites are almost all in the public domain today – thus providing an excellent starting point for digitization. Other materials can provide other categories of background: (Hall 1913), for example, describes the textual traditions for all major classical authors as it was understood in the early twentieth century (and thus as it appears in most public domain editions. Authors and works that have appeared as separate editions also have standard names: once we associate Marcus Tullius Cicero, M. Tullius Cicero, and Cicero, for example, with the canonical name authority form “Cicero, Marcus Tullius,” we can automatically search standard library catalogues. Online texts generally provide one citation scheme. Some authors, however, have multiple citation schemes and we need to manage them all if we are to exploit the full range of citations. These should be included when the electronic editions are created, with alternate citation schemes added to existing texts as image books with the alternate citations become available.

- Lexicography: We want to be able to identify not only particular forms and dictionary entries but the distinct senses of particular words in particular passages. Parallel corpora, with source texts in one language aligned with translations in one or more languages, have allowed machine translation to make substantial progress in recent years. The machine translation systems can look for statistical associations between words in the two languages to identify probable translation equivalents for particular words in particular passages. Machine readable dictionaries remain crucial tools for machines as well as for human readers.⁸³ Online lexica not only provide reading support but provide a foundation for semantic analysis through comparison of dictionary definitions and an open inventory of documented senses. LSJ 9 and Lewis and Short,⁸⁴ augmented by more specialized lexica, provide a reasonable starting point for an electronic infrastructure.
- Syntax: We also want to be able to identify the syntactic relations within a sentence – at the simplest, answering the question “what does this word depend on and what is its function?” Generated accurate parse trees for complete sentences is difficult in any case and increasingly difficult the larger and more complex the sentence. Nevertheless, even if the complete sentence parse is not correct, enough individual word-to-word relations are usually correct to detect patterns such as which nouns go with which adjectives, what cases a verb takes etc. Grammars are the logical starting point for syntactic data: we have thus digitized the extensive Kühner-Gerth Greek Grammar,⁸⁵ as well as the shorter

⁸³ For more on machine translation and WSD see (Smith 2006), (Marcu 2005), and (Carpuat 2005).

⁸⁴ (Liddell et al. 1940), (Andrews et al. 1879).

⁸⁵ (Kühner et al. 1890).

Smyth⁸⁶ and Allen and Greenough⁸⁷ grammars for Greek and Latin. Highly inflected languages store much of their syntactic information in word forms that less heavily inflected languages may express in word order. Greek and Latin lexica thus contain much – and arguably more – syntactic information than conventional grammars, since the constructions associated with individual words may be key to determining the correct parses for a sentence.

- Specialized reference materials: Larger works may contain specialized glossaries on particular topics. (Hall 1913) contains a very useful glossary that explains the Latin names for manuscripts in many editions; (Smyth 1920) contains a glossary of rhetorical terms. Specialized lexica cover the language of particular authors, such as Slater's Pindar lexicon.⁸⁸ Once again, the Smith dictionary series provided us with a foundational resource on which to build: the two-volume *Dictionary of Greek and Roman Antiquities* (Smith, Wayte et al. 1890) contains 3,400 entries and (as of this writing) 25,000 extracted citations covering law, architecture, religion, rhetoric and other aspects of life.
- Events: One can easily enter a philosophical funk trying to define what is and is not an event, but modern timelines and ancient chronologies show us what others chose to identify as significant events and provide us with an objective record of what others have chosen to label and recall.⁸⁹

The role of the editor in a digital world

The digital world makes possible a new kind of editor: the corpus editor occupies a middle ground between the algorithm heavy, knowledge light approaches of computer science and the wholly manual practices of traditional editing. The corpus editor works with thematically coherent bodies of text that are too big to be processed and checked by hand and that therefore demand automated methods. The corpus editor combines knowledge bases and automated methods to apply automated markup and/or extract information. The corpus editor cannot check every automated decision but is able to document both how the automated decisions were made and to provide statistical measures for the accuracy of those decisions.⁹⁰

The role of the traditional editor also changes in an electronic environment. The traditional editor becomes responsible for preparing documents for use not only by people but by machines. The ePhilologist reviews a high percentage – and ideally all – of the automated decisions that link a particular text to knowledge sources such as those listed above: the editor manages the automated processes and reviews the results. The editor checks the morphological analyses and parse trees, comments on passages where

⁸⁶ (Smyth 1920).

⁸⁷ (Allen et al. 1904).

⁸⁸ (Slater 1969).

⁸⁹ The use of HEML (Historical Event Markup Language) could be applicable in this area; see (Robertson 2006).

⁹⁰ For more on the role of corpus editors, see (Crane 2000).

the identification of a person or place is ambiguous, etc. The edited documents in the digital library provide crucial training sets that improve the performance of automated methods generally: thus, careful work on a few lives of Plutarch should improve results on the other lives and on similar Greek prose generally.

Cultural Informatics

Digital culture already dominates serious intellectual life, even if its dominance still subordinates itself to the superficial – and, to a classicist, quite recent – forms of print culture. The previous section described one partial survey of what form classics might take as a digital culture matures and intellectual practice begins to exploit this digital world for its own strengths. The examples given reflect substantive work with existing technologies applied to questions common to all students of historical languages. All of the examples above either are, or could become, general services.⁹¹ Nevertheless, they constitute a few first steps in a much larger process.

Much of the above work was possible because the National Science Foundation and the National Endowment for the Humanities collaborated on the Digital Library Initiative Phase II, a program which supported a range of humanities projects. We cannot expect such levels of support in the future.⁹² If we are to move forward as a field, we must use what we have learned from what worked and what did not work in the past to develop a strategy to help us move forward in the future. Classics may or may not pursue the particular directions suggested in the previous section, but passively drifting along a broader current of academic practice is a dangerous course. The Mellon Foundation and American Council on Learned Societies recently funded a “Commission on Cyberinfrastructure for Humanities and Social Sciences.”⁹³ A PhD in English (John Unsworth) chaired the commission, which included five humanists, including another person from English literature (Jerome McGann), an American historian (Roy Rosenzweig), an art historian (Sarah Fraser), and the director of an archaeological research collection (Bruce Zuckerman). The draft report available in May 2006 makes cursory mention of classics. Classicists cannot expect colleagues who work primarily in English and with relatively recent sources to anticipate the problems of working with historical languages. Classics – and all disciplines which draw upon languages of the past – must tirelessly engage on larger conversations and be prepared to defend the significance of language.

One effective solution is the creation of a new area of informatics designed to bridge the gap between a discipline and current research in computer science – a demanding task, if performed well, because it requires a command of emerging, as well as established, issues in two radically different disciplines. The field of biology, confronted with overwhelming amounts of raw data, produced the field of bioinformatics, thus creating an intellectual space, primarily grounded in biology, to connect research in computer science with biology research.

⁹¹ For examples of potential services, please see (Patton 2004) and (Crane et. al. 2006b).

⁹² For a discussion for the future of digital library funding, see (Griffin 2006).

⁹³ <http://www.acls.org/cyberinfrastructure/cyber.htm>.

Classics probably cannot command a hundredth part of the resources on which biological research depends. We cannot call forth a major new discipline with the funding to attract the attention of grant-driven computer scientists. Nevertheless, we can accomplish a great deal.

- All philological inquiry, whether classical or otherwise, is now a special case of corpus linguistics. Its foundational tools should come increasingly from computational linguistics, with human and automated analysis. Vague statements such as “typical of Greek prose,” “common in early Greek” etc. must give way to dynamically generated measurements of well-mapped corpora. Human judgment must work draw upon and work in conjunction with documented mathematically grounded models. The salaries which support Classics faculty are the one resource which we, as a field, collectively allocate. As at least some, if not most, members of the field begin to see themselves as computational linguists with a particular focus on Latin and/or Greek, we will soon mobilize over a long period of time far more intellectual capital than the most generous grants could provide for limited periods.
- We need to rethink what we study. Tasks which we as human readers take for granted often demand substantial analysis when we transfer them to automated systems. Classicists cannot manually fit all 91,000,000 words in the current TLG into parse trees. Some tasks, such as concordance generation or even more sophisticated problems such as morphological analysis, can follow well-defined results: e.g., display all possible instances of the Latin verb *facio* in the *Catilinarians* of Cicero. We soon reach problems for which rule sets provide much less accurate results: is a given instance of the form *faciam* a subjunctive or future? Which “Alexander” does a particular passage cite? Which accusative noun is the subject of the infinitive and which the object? We need a foundational work on the problem of resolving ambiguities, producing the best possible results and providing accurate information as to the accuracy of automated results. We need to take a step and work on the tools on which research will rely.
- We must distinguish programming from computer science. We will need quite a bit of advanced programming, even if we are only gluing together tools developed by our colleagues in computational linguistics. Nevertheless, we must separate analysis of our methods from the code by which we test them. We need the patience to evaluate multiple methods to solve the same problem and to produce results from which others can learn – a patience that will become more common as we develop a community of research. We also need to consider our skills: crucial as programming may be, philologists who wish to draw effectively upon the emerging tools of our world must become familiar with linear algebra and probability.
- ePhilology is part of a larger, cultural informatics. ePhilology represents one particular approach to a comprehensive analysis of earlier culture: we may center our attention on words, but our questions will soon lead us to the evidence of

material culture. Classics may be big enough to sustain its own classical informatics, but we would be much better served by contributing to a larger cultural informatics. We should aggressively establish alliances with partners with similar needs and limit, as much as possible, ourselves to those problems which only classicists can address. We have developed our own morphological analyzers, syntactic analyzers and named entity recognition systems, but it would be much better for us to concentrate on the databases of stems and endings, the grammar, and the knowledge bases of people, places etc. Our natural collaborators include not only all of those working with historical languages but also those struggling to analyze the thousands of languages spoken in our contemporary world. Where cultural informatics would embrace all sources of information – natural language, relational databases, images, GIS, 2D and 3D models, simulations – ePhilology implies a focus upon linguistic sources.

- We need to identify what structures we need to institutionalize: A generation ago, classicists could get jobs, tenure and promotion at leading institutions as editors and authors of scholarly commentaries. Almost all classics faculty under the age of fifty in US departments have, however, made their careers by producing articles and monographs, with far less emphasis on editing and work on the intellectual infrastructure of the field. A generation of ePhilologists may emerge to play prominent roles in our departments as the field realizes that we are not just copying print into digital form but creating a wholly new, qualitatively distinct infrastructure. The changes before us may exceed those spawned by movable print and may be more comparable to the invention of writing itself. We also need new libraries to help us maintain into the future the resources that we create. Libraries will need to develop new skills to manage digital libraries and new ways to use their acquisition budgets to support the creation of content with the structure and the right regimes needed by humanists.⁹⁴ We may need new departmental and research structures – combinations of Classics and computer science may become common, to the benefit of both fields. We need to establish relationships with major commercial entities such as Google, Yahoo and Microsoft, if these continue to evolve into the public libraries of the twenty-first century and provide us with new channels to society as a whole.

Conclusion

Some emerging technologies could, if applied to classics and to other philological disciplines, have a swift and dramatic impact upon the questions that we pursue: machine translation, parallel text analysis, named entity identification, syntactic analysis, cross language information retrieval and a range of text mining methods are well suited to a range of needs. The impact of digital technology will, however, be far broader and more pervasive than any particular tools we can deploy in the immediate future. The future of classics depends less upon particular tools than upon an emerging digital environment that integrates an increasing number of tools together into a dynamic world,

⁹⁴For more on the needs of new library services and infrastructures, see (Dempsey 2006).

constantly evolving to answer our questions and support the life of the mind. From the nineteenth century through the twentieth, we were able to take our scholarly infrastructure for granted: we had our publishers and libraries, our editions, commentaries, lexica, journals, monographs, and encyclopedias. We now have the merging of print, broadcast media and gaming, new commercial entities planning universal access to a better library than the wealthiest academic institution on earth could provide to its faculty; we have new forms of intellectual production such as blogs and wikis; we have ontologies and knowledge bases at the core of reference materials; we have a world of dynamic information – books that read and learn from each other and from their human readers. The challenge now – and it is perhaps the greatest challenge classicists have faced since they found themselves pushed out of the center of the academy – is to shape this world and negotiate a new place for classical studies within it.

Bibliography

- Allen, J. H., J. E. Greenough, et al. (1904). Allen and Greenough's New Latin grammar for schools and colleges, founded on comparative grammar. Boston London, Ginn & company.
- Andrews, E. A., W. Freund, et al. (1879). A Latin dictionary founded on Andrews' edition of Freund's Latin dictionary. Oxford ; New York, Clarendon Press.
- Audenaert, N, et. al. "Integrating Collections at the Cervantes Project." JCDL 2005, pp. 287-8.
- Barzilay, Regina and Noemie Elhadad. (2003). "Sentence alignment for monolingual comparable corpora". In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003), pp. 25–32.
- Beagrie, N. (2005). "Plenty of room at the bottom? Personal digital libraries and collections." D-Lib Magazine, June, 11(6), <http://dlib.anu.edu.au/dlib/june05/beagrie/06beagrie.html>
- Bearman D. and J. Trant. (2005). "Converting scanned images of the print history of the world to knowledge: a reference model and research strategy." RDLP, 8 (5), <http://www.elbib.ru/index.phtml?page=elbib/eng/journal/2005/part5/BT>
- Bekaert, J. and H. Van de Sompel. (2006). "Augmenting interoperability across scholarly repositories." <http://eprints.rclis.org/archive/00006924/>
- Berkowitz, L. and K. A. Squitier (1990). Thesaurus Linguae Graecae Canon of Greek Authors and Works. New York, Oxford University Press.
- Bia, A., C., I.Garrigós, and J. Gómez. (2004). "Personalizing digital libraries at design time: The Miguel de Cervantes digital library case study." Web Engineering, pp. 225-9.

- Biblioteca Teubneriana Latina, BTL-3 (Turnhout: Brepols; Munich: K. G. Saur, 2004).
- Blackwell, C. and N. Smith (2005) "A Guide to version 1.1 of the Classical Text Services Protocol." Digital incunabula: a CHS site devoted to the cultivation of digital arts and letters. <http://chs75.harvard.edu/projects/diginc/techpub/cts-overview>
- Bodard, G. (2006). "Inscriptions of Aphrodisias: Paradigm of an electronic publication." CLiP 2006, <http://www.cch.kcl.ac.uk/clip2006/content/abstracts/paper33.html>
- Bowen, J. P. and S. F. Fantoni. (2004). "Personalization and the Web from a Museum Perspective." Musems and the Web 2004, <http://www.archimuse.com/mw2004/papers/bowen/bowen.html>www.archimuse.com/mw2004/papers/bowen/bowen.html
- Brusilovsky, P, S. Sosnovksy, and O. Shcherbinina. (2005). "User modeling in a distributed E-learning architecture." User Modeling 2005, LNCS 3538, pp. 387-391.
- Carpuat, M. and D. Wu. (2005). "Word sense disambiguation vs. statistical machine translation." Proceedings of the Association for Computational Linguistics 2005, pp. 387-94.
- Carrera, F. (2005). "Making history: An emergent system for the systematic accrual of transcriptions of historic manuscripts." Eighth International Conference on Document Analysis and Recognition (ICDAR'05), pp. 543-9.
- Chklovski, T. and Gil, Y. 2005. "Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors." In Proceedings of the 3rd international Conference on Knowledge Capture, K-CAP '05, pp.35-42.
- Church, Kenneth, and Patrick Hanks (1989). "Word association norms, mutual information, and lexicography," ACL 27, pp. 76-83.
- Crane, G. and Rydberg-Cox, J. A. (2000). "New technology and new roles: the need for "corpus editors". In Proceedings of the Fifth ACM Conference on Digital Libraries, pp. 252-253.
- Crane, G. (2004). "Classics and the computer: an end of the history," in A Companion to the Digital Humanities, edited by Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell Publishing, 2004.
- Crane, G. and A. Jones. (2005). "The Perseus American Collection 1.0." www.perseus.tufts.edu/~gcrane/americancoll.12.2005.pdf

- Crane, G. and A. Jones. (2006). "The challenge of Virginia Banks: An evaluation of named entity analysis in a 19th Century newspaper collection." Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, pp. 31-40.
- Crane, Gregory and Alison Jones. (2006). "Text, information, knowledge and the evolving record of humanity." D-Lib Magazine, March, 12 (3), <http://purl.pt/302/1/dlib/march06/jones/03jones.html>.
- Crane, G. (2005a). "No book is an island: designing electronic primary sources and reference works for the humanities," in H. van Oostendorp, Leen Breure, Andrew Dillon, eds, Creation, Use, and Deployment of Digital Information, (Erlbaum 2005), pp. 11-26.
- Crane, G. (2005b). "Reading in the age of Google : Contemplating the future with books That talk to one another," in Humanities, September/October, 26 (5), <http://www.neh.gov/news/humanities/2005-09/readingintheage.html>.
- Crane, G. et. al. (2006). " Beyond digital incunabula: Modeling the next generation of digital libraries." ECDL 2006, pp. 353-66.
- Crane, G. et. al. (2006b) "Services make the repository." Paper presented at JCDL 2006 Workshop, Digital Curation and Trusted Repositories, <http://www.ils.unc.edu/tibbo/JCDL2006/Jones-JCDLWorkshop2006l.pdf>
- de Jong, F. et al.. (2005). "Temporal language models for the disclosure of historical text", XVIth International Conference of the Association for History and Computing, 2005.
- Dekhytar, A. et, al. (2005). "Support for XML markup of image based electronic editions." International Journal on Digital Libraries, pp.55-69.
- Dempsey, L. (2006). "The (digital) library environment: Ten years after." Ariadne, <http://www.ariadne.ac.uk/issue46/dempsey/>.
- Derrida, J. (1972). "La Pharmacie de Platon," in: La Dissemination (Paris: Éditions du Seuil), pp. 69-196.
- Doerr, M., J. Hunter and C. Lagoze. (2003). "Towards a core ontology for information integration." Journal of Digital Information, 4 (1).
- Dolog, P., et. al. (2004), "The personal reader: Personalizing and enriching learning Resources Using Semantic Web Technologies." Proceedings of the 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, pp. 85-94.

- Eide, O. and C. E. Ore. (2006). "TEI, CIDOC-CRM and a Possible Interface between the Two." Proceedings of the ALLC-AHC 2006.
- Egan, G. (2005). "Impalpable hits: Indeterminacy in the searching of tagged Shakespearian Texts." Paper Delivered on 17 March at the 33rd Annual Meeting of the Shakespeare Association in America, in Bermuda.
<http://magpie.lboro.ac.uk/dspace/handle/2134/1294>
- Esposito, F., et. al. (2005). "Intelligent document processing." Proceedings of Eighth International Conference on Document Analysis and Recognition, pp. 1100-1104.
- Felluga, D. F. (2005). "Addressed to the NINES: the Victorian archive and the disappearance of the book." Victorian Studies, 48 (2), pp. 305-319.
- Garrett, J. (2006). "KWIC and dirty? Human cognition and the claims of full text searching." Journal of Electronic Publishing, 9 (1),
<http://www.hti.umich.edu/jjep/>.
- Gemmell, J., G. Bell, and R. Lueder. (2006). "MyLifeBits: A personal database for everything." Communications of the ACM, January, 49 (1), pp. 88-95.
- Gerlach, A. E. and N. Fuhr. (2006). "Generating search term variants for text collections with historic spellings." ECIR 2006, pp. 49-60.
- Gilardoni, L. et. al. (2005). "Machine learning for the Semantic Web: Putting the user into the cycle." Published in the Proceedings of the Dagstuhl Seminar on Machine Learning for the Semantic Web, 13-18 February 2005, Dagstuhl, Germany,
www.quinary.com/pagine/downloads/files/Resources/QuinaryDagstuhl.pdf
- Griffin, S. (2005). "Funding for digital libraries: Past and present." D-Lib Magazine, 11 (7/8), <http://www.dlib.org/dlib/july05/griffin/07griffin.html>
- Hall, F. W. (1913). A companion to classical texts. Oxford: Clarendon press.
- Hardwick, L. (2000). "Electrifying the canon: The impact of computing on classical studies." Computers and the Humanities, 34, pp. 279-95.
- Hinman, C. (1968). The First Folio of Shakespeare: The Norton Facsimile. New York, W. W. Norton.
- Hoekstra, R. (2005). "Integrating structured and unstructured searching in historical sources. In Proceedings of the XVI International Conference of the Association for History and Computing, pp. 149-54.
- Hornblower, S. and A. Spawforth (1996). The Oxford classical dictionary. New York, Oxford University Press.

- Ignat, C., et. al. (2005). "A tool set for the quick and efficient exploration of large document collections." Proceedings of the 27th Annual ESARDA Meeting.
- Justeson, John S., and Slava M. Katz (1995). "Technical terminology: some linguistic properties and an algorithm for identification in text," Natural Language Engineering 1, pp. 9-27.
- Kavcic, Alenka. (2004). "Fuzzy user modeling for adaptation in educational hypermedia." IEEE Transactions on Systems, Man and Cybernetics, Part C, November, 34 (4), pp. 439-449.
- Kelly, K. (2006). "Scan This Book!" New York Times Magazine.
<http://www.nytimes.com/2006/05/14/14publishing.html>
- Kim, S. and E. A. Fox. (2004). " Interest-based user grouping model for collaborative filtering in digital libraries." ICADL, pp. 533-42.
- Kirschenbaum, M. (2006). "The NORA Project: Text mining and literary interpretation." Digital Humanities 2006, pp. 255-6.
- Kolbitsch, J. and H. Maurer. (2005). "Community building around encyclopaedic knowledge." Journal of Computing and Information Technology.
- Kruk, S.R., S. Decker, and L. Zieborak. (2005). "Adding Semantic Web technologies to digital libraries." DEXA 2005, LNCS 3588, pp. 716-725.
- Kühner, R., F. Blass, et al. (1890). Ausführliche grammatik der griechischen sprache. Hannover, Hahnsche buchhandlung.
- Latousek, R. (2001). "Fifty years of classical computing: A progress report." CALICO Journal, 18 (2), pp. 211-22.
- Liddell, H. G., R. Scott, et al. (1940). A Greek-English lexicon. Oxford, The Clarendon Press.
- Linden, G., B. Smith, J. York. (2003). " Amazon.com recommendations: Item-to-item collaborative filtering." Internet Computing, 7 (1), 76-80.
- MacColl, J. (2006). "Google challenges for academic libraries." Ariadne, 46,
<http://www.ariadne.ac.uk/issue46/maccoll/>
- Marcu, D. and K. Knight. (2005). "Machine translation in the year 2004," in Proceedings of Acoustics, Speech and Signal Proceedings (ICASSP 2005), Volume 5, pp. 965-8.
- McManus, B. F. and C.A. Rubino. (2003). "Classics and Internet technology." American Journal of Philology, 124 (4), pp. 601-8.

- Mihalcea, Rada and Michel Simard. (2005). "Parallel texts." Natural Language Engineering, September, 11 (3), pp. 239-46.
- Mirzaee, V., et. al. "Computational representation of semantics in historical documents." Proceedings of AHC 2005.
- Nagy, G. and D. Lopresti. (2006). "Interactive document processing and digital libraries." Proceedings of the Second International Conference on Document Images, Analysis for Libraries (DIAL 2006), pp. 2-11.
- Nagypal, G., et. al. (2005). "Applying the Semantic Web: The VICODI experience in creating visual contextualization for history." Literary and Linguistic Computing, 20 (3), pp. 327-349.
- Nagypal, G. (2004). "Creating an application-level ontology for the complex domain of history: mission impossible?" In Proceedings of Lernen—Wissensentdeckung—Adaptivitaät (LWA 2004), FGWM 2004 Workshop, Berlin, Germany, pp. 287–94. http://lwa.informatik.hu-berlin.de/proceedings/LWA04_FGWM.pdf (accessed 12 April 2005).
- Navigli, R. and P. Velardi. (2005.) "Structural semantic interconnections: A knowledge-based approach to word sense disambiguation." IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1075-86.
- Niederée, C., et. al. (2004). "A multi-dimensional, unified user model for cross-system personalization." In Proceedings of Advanced Visual Interfaces International Working Conference (AVI 2004) -Workshop on Environments for Personalized Information Access, Italy, pp. 34-54.
- Packard, D. W. (1973). "Computer-assisted morphological analysis of ancient Greek." Proceedings of the 5th Conference on Computational Linguistics, Pisa, Italy, pp. 343-55.
- Patton, M. S. and D. M. Mimno. (2004). "Services for a customizable authority linking environment." In Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, p. 420.
- Pierazzo, E. "Just different layers? Stylesheets and digital edition methodology." Digital Humanities 2006.
- Plaisant, C., J. Rose, et al. (2006). Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces. Joint Conference on Digital Libraries, Chapel Hill, NC, ACM Press.
- Porter, D., et. al. (2006). "Creating CTS collections." Digital Humanities 2006, pp. 269-274.

- Riva, M. and V. Zafrin. (2005). "Extending the text: Digital editions and the hypertextual paradigm." Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia, pp. 205-207.
- Robertson, B. (2006). "Visualizing An historical Semantic Web with HEML." Proceedings of the WWW 2006, pp. 1051-2.
- Rosenzweig, Roy. (2006). "Can history be open source: Wikipedia and the future of the past." *Journal of American History*, 93 (1), pp. 37-46.
- Rouane, K. and C. Frasson and M. Kaltenbach. (2003). "Reading for understanding: A framework for Advanced Reading Support." Proceedings of the 3rd IEEE International Conference on Advanced Learning Technologies, pp. 394-5.
- Ruhleder, K.. (1995). "Reconstructing artifacts, reconstructing work: From textual edition to on-line databank." *Science, Technology, & Human Values*, 20 (1), Winter, pp. 39-64.
- Russell, J. (2003). "Making it personal: Information that adapts to the reader." SIGDOC '03: Proceedings of the 21st Annual International Conference on Documentation, pp. 160–166.
- Sankar, K.P., et. al. (2006). "Digitizing a million books: Challenges for document analysis." Document Analysis Systems VII, 7th International Workshop, DAS 2006, pp. 225-36.
- Schmidt, D. and T. Wyeld. (2005). "A novel user interface for online literary documents." Canberra, Australia. November 23-25, pp. 1-4.
- Shoemaker, R. (2005). "Digital London: Creating a searchable web of interlinked resources on eighteenth century London." Program: Electronic Library and Information Systems, 39(4):297–311.
- Shamos, M. I. (2005). "Machines as readers: a solution to the copyright problem." *J. Zhejiang Univ. Science* 6A, 11, pp. 1179-1187.
- Siemens, R. "Knowledge management and textual cultures? Work toward the Renaissance English Knowledgebase (REKn) and its professional reading environment." CASTA 2006.
- Slater, W. J. (1969). Lexicon to Pindar. Berlin, de Gruyter.
- Smeaton, A. F. and J. Callan. (2005). "Personalisation and recommender systems in digital libraries." Int. J. Digit Lib, 5: 299-308.

- Smith, D. A. (2006). "Debabelizing libraries: Machine translation by and for digital collections." D-Lib Magazine, March, 12 (3), <http://www.dlib.org/dlib/march06/smith/03smith.html>.
- Smith, D. A. and G. Crane. (2001). "Disambiguating geographic names in a historical digital library." Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'01), Lecture Notes in Computer Science, pp. 127–136.
- Smith, W. (1854). Dictionary of Greek and Roman geography. Boston, Little Brown & co.
- Smith, W. (1873). A dictionary of Greek and Roman biography and mythology. London, J. Murray.
- Smith, W., W. Wayte, et al. (1890). A dictionary of Greek and Roman antiquities. London, J. Murray.
- Smyth, H. W. (1920). A Greek grammar for colleges. New York, Cincinnati [etc.], American Book Company.
- Spencer, M. and C. Howe. (2004). "Collating texts using progressive multiple alignment," Computers and the Humanities, August, 38 (3), pp. 253-70.
- Tennant, R. (2005). "The Open Content Alliance." Library Journal, December 15, 2005, <http://www.libraryjournal.com/article/CA6289918.html>
- Terras, M. (2005). "Reading the readers: Modelling complex humanities processes to build cognitive systems." Literary and Linguistic Computing, 20 (1), pp. 41-59.
- Thatcher, S. G. (2006). "Fair use in theory and practice: Reflections on its history and the Google Case." Journal of Scholarly Publishing, 37 (3): 215-229.
- Travis, H. (2005). "Building universal digital libraries: An agenda For copyright reform." Forthcoming, Pepperdine Law Review, available at <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=793585>.
- Unsworth, J. (2003). "The crisis in scholarly publishing in the humanities." ARL Bimonthly Report 228, <http://www.arl.org/newsltr/228/crisis.html>.
- Wang, C.Y. and G.D. Chen. (2004). "Extending e-books with annotation, online support and assessment mechanisms to increase efficiency of learning." SIGCSE Bulletin, 36 (3), pp. 132-136.

- Xiang, X. and J. Unsworth. (2006). "Connecting text mining and natural language processing in a humanistic context" Digital Humanities 2006, The Sorbonne University, July 6, 2006.
- Volkel, M, et.al. (2006). "Semantic Wikipedia." *WWW 2006*, pp. 585-594.
- Willinsky, J. (2005). *The access principle: The case for open access to research scholarship*. Cambridge, MA: MIT Press.
- Willinsky, J. (2003). "Opening access: Reading (research) in the age of information." In C. M. Fairbanks, J. Worthy, B. Maloch, J. V. Hoffman, & D. L. Schallert, (Eds.), 51st National Reading Conference Yearbook, Oak Creek, WI: National Reading Conference, pp. 32-46.
- Witte, Rene. (2005). "Engineering a semantic desktop for building historians and architects". *SemDesk 2005 Workshop Proceedings*.