Tools for Thinking: ePhilology and Cyberinfrastructure

Gregory Crane, Alison Babeu, David Bamman, Lisa Cerrato, Rashmi Singhal

philology brings back to life the words of languages no longer spoken. While literally "the love of language," philology includes not only linguistics but philosophy, history, literary criticism, the history of science and technology, political science, economics, art, archaeology, and every other discipline relevant to the world that these texts describe. Of course, philology must, in its fullest form, engage fully with the material record: museum collections and archeological excavations not only serve to illustrate topics within the text but also provide independent windows onto the past from which we may survey views very different from those we glimpse in the texts alone. Philology is thus not just about text; it is about the world that produced our surviving textual sources and about the tangible impact that these texts have had upon the worlds that read them.¹

Few of us manage to be philologists in this broad sense. We cannot, with the tools of print technology, cover enough intellectual ground. Even if we set aside, for the moment, the problem of working with material culture, and consider only the challenges of textual materials easily represented in print form, our limitations are severe. As Solon points out in *The History of Herodotus*, there are only about 30,000 days in a human life—at a book a day, we would need 30 generations to read through even a moderate collection of a million books and 10,000 years to cover the 10 million-or-so unique items in the Harvard Library system.

The barriers are not simply quantitative. Few of us will ever be able to finish a cursory reading of 10 books, however thin, if these

¹ For some further exploration of the new challenges of e-philology, see G. Crane, D. Bamman, and A. Babeu. 2008. ePhilology: When the Books Talk to Their Readers. In *A Companion to Digital Literary Studies*, edited by Ray Siemens and Susan Schreibman. New York: Blackwell Publishing, 29–64.

books contain untranslated poems in 10 different languages. Classical philologists must have expertise in Greek and Latin and an ability to work with scholarship in English, French, German, and Italian. If, however, we wish to explore broader topics that cut across multiple cultures, e.g., the impact of Genghis Khan and his successors or the rise of Christianity and Islam, then we soon confront sources in far more languages than most scholars can expect to master. And indeed, in many cases mastery may not be an issue: scholars are still rapidly expanding our ability to understand languages such as Sumerian and Mayan. In cases such as classical Greek, Sanskrit, and Chinese, by contrast, so much information survives that we must remain students for our entire lives.

The great challenge for the rising generation of scholars is to build a digital infrastructure with which to expand our intellectual range.² We seek to advance two effects already enabled by the digital infrastructure at hand. On the one hand, we are extending the intellectual range of individual scholars, enabling them to pursue topics that require analysis of more primary sources or more linguistic materials than was feasible with print. Mark Schiefsky's work with Archimedes illustrates how scholars were able to explore a broad historical topic (in this case, the history of mechanics) with greater rigor than would have been possible in print—assuming they would have undertaken such an ambitious project at all. At the same time, we want to increase the complementary effect and further extend the audiences that the products of particular cultures can reach. Machine translation is one technology that aims to advance this goal, but even the simple translation-support systems already provided in environments such as the Perseus Digital Library have for years made foreign language texts intellectually more accessible to students than print resources alone.

We can already see new classes of research project taking shape. Thus, we could, with existing technology, build collections and services in which we could study the influence of Plato across a wide range of cultures, including not only every written language from the history of Europe but Arabic and Persian as well. Multilingual named entity identification systems would scan these corpora for references to Plato, for translations of his works, and for quotations of particular passages.³ Text-mining systems would summarize pat-

² A good overview of the challenges of building a digital infrastructure in the humanities, as well as a survey of much of the recent literature on the topic, can be found in D. Green and M. Roy. 2008. Things to Do While Waiting for the Future to Happen: Building a Cyberinfrastructure for the Liberal Arts. *EDUCAUSE Review* 43(4), available at http://connect.educause.edu/display/46969. For a look at some of the specific challenges for building a cyberinfrastructure in classics, see D. Pritchard. 2008. Working Papers, Open Access, and Cyber-infrastructure in Classical Studies. *Literary and Linguistic Computing* 23(2): 149–162.

³ Multilingual systems for named entity recognition is an area of research that is growing rapidly. For some interesting recent work in this area, see C. Silberer, et al. 2008. Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 3230-3237. Available at http://www.lrec-conf.org/proceedings/lrec2008/pdf/816_paper.pdf.

terns of thought associated with Plato and his works.⁴ Word- and phrase-translation systems would allow us to extract the meanings of these key phrases in languages with which we are not familiar. We could even begin to align patterns in different languages, perhaps revealing that discourse about Plato in ninth-century Arabic is more closely related to that in Persian than to that in nineteenth-century German—or perhaps not.

These automated processes are only a starting point. Like the infrastructure of modern athletic training, our intellectual infrastructure only allows us to use our limited cognitive resources to greater effect. Our customization and personalization systems would use models of our educational background and immediate purposes⁵ to provide us with the briefing materials necessary to begin evaluating what we see: pointers to translations into languages with which we are familiar (e.g., from Persian into French), automatically generated lists of new words and concepts in sources where we have studied the documents, pre-existing encyclopedia entries, and automatically generated key phrases in recent scholarship about people, places, organizations and readily identified topics (e.g., Plato's *Republic*).

We already have the algorithms, and Google—or the Google partner libraries with noncommercial rights to books digitized from their collections—have the collections⁶ that would open new areas of research that become possible only when we can automatically analyze collections far too big and far too heterogeneous for any human brain.

Consider one concrete example. In 2010, 2,500 years will have passed since the Greeks confronted an army from the Persian Empire on the plains of Marathon. After 10 years of training, a junior classicist might have extensive, but hardly exhaustive, knowledge of the scholarship surrounding Herodotus's accounts of the Persian Wars in the early fifth century or the major Greek sources about Alexander's invasion of Persia a century and a half later. With a good deal of effort, the junior classicist could develop an undergraduate survey course about Greek and Persian relations, as seen from Greek and Latin sources. One scholar suggested in private correspondence that 95 percent of the research on Alexander the Great involves scholars

⁴ The potential of text mining for humanities texts has been explored in recent years by various researchers. For some recent work, see A. Don, et al. 2007. Discovering Interesting Usage Patterns in Text Collections: Integrating Text Mining With Visualization. *Proceedings of the Sixteenth ACM Conference on Information Knowledge Management*, 213–222.

⁵ Customization and personalization systems that utilize user models to adapt information to the reader's needs have been explored by many researchers, such as F. Ahmad, et al. 2007. Towards Automatic Conceptual Personalization Tools. *JCDL* 2007: Proceedings of the 7th ACM/IEEE CS-Joint Conference on Digital Libraries, 452-461; and E. Frias Martinez, et al. 2006. Automated User Modeling For Personalized Digital Libraries. *International Journal of Information Management* 26(3): 234–248.

⁶ An article by Grogg and Ashmore explores how Google partner libraries vary in what they are doing with their digital copies. See J. E. Grogg and B. Ashmore. 2007. Google Book Search Libraries and their Digital Copies. Searcher 15(4). Available at http://www.infotoday.com/searcher/apr07/Grogg_Ashmore. shtml.

who do not know a word of Old Persian and have no substantive knowledge of Iranian civilization. Whether or not this admittedly subjective estimate is accurate, the multiethnic and multilingual nature of the Persian Empire has split the subject into small, isolated communities.

There are two problems. First, scholars simply do not have physical access to the sources illustrating Iranian civilization. Second, even if they did have physical access, few can read Farsi, or even put their hands on the background materials needed to contextualize sources about Iranian civilization. Even if the information is available in our existing library collections, scholars are not synthesizing that information. Scholars have adapted their work to the limits of what they can accomplish. All responsible scholars of Alexander would welcome an infrastructure that would allow them to understand the subject as widely as possible. Existing scholarship reflects harsh compromises, as scholars learned what their cognitive resources could accomplish in the tools of print. We need a digital infrastructure that can assemble primary and secondary sources now scattered throughout specialized publications and then provide the background information that each scholar needs to carry on his or her work.

This leads us to the second major advance of the emerging digital infrastructure: if we can change the intellectual range of individual human thinkers, we can also increase the audience for individual products of human culture. By automatically linking inflected words in a text to linguistic analyses and dictionary entries, we have already allowed readers to spend more time thinking about the text than was possible as they flipped through print dictionaries. Reading-support tools allow readers to understand linguistic sources at an earlier stage of their training and to ask questions, no matter how advanced their knowledge, that were not feasible in print. In effect, as we provide more and more sophisticated reading support, we extend the intellectual reach of complex cultural productions.

More than 2,000 years ago, Plato's Socrates questioned the value of written information if it is not converted to active knowledge in a human brain. If we in the humanities had to choose, many of us would agree that it is more important to help the current body of ideas about antiquity play a more vibrant role in human society than to produce new ideas intellectually accessible to their established audiences (i.e., those with years of training and with professional access to libraries that pay for digital subscriptions), but we might find ourselves hard-pressed to make a decision. Some, perhaps most, of us who are professional humanists believe that we have a primary obligation to make the human record play the most dynamic

⁷ Reading support tools that help readers more effectively mine their way through digital text is a growing area of research. For some interesting examples, see E. H. Chi, et al. 2007. ScentIndex and ScentHighlights: Productive Reading Techniques for Conceptually Reorganizing Subject Indexes and Highlighting Passages. *Information Visualization* 6(1): 32–47; and C. Faire and N. Vincent. 2007. Document Image Analysis for Active Reading. *Proceedings of the International Workshop on Semantically Aware Document Processing and Indexing*, Montpellier, France, May 21-22, 2007, pp. 7–14.

role possible in the intellectual life of humanity. The two go hand in hand: the more intellectual activity around a topic, the more intellectual labor available. Gutenberg printed Latin bibles. Martin Luther, William Tyndale, and others, building on the technology of print, translated the Bible and fostered intellectual communities that had not previously existed. They changed the world.

It would be easy enough to explore in 2010 the Greco-Roman view of the Battle of Marathon 2,500 years before. We would, however, rather broaden the discussion and engage Iranian scholars to provide their perspectives on the Achaemenid Empire. Ideally, the major sources, including both the textual and material record, would be freely available in digital form, with reading support and other background information in place. Those of us who have dedicated our lives to the study of the Greco-Roman world would welcome the tools whereby we could understand, as deeply as possible, how the fifth-century BCE appears to those who see the Persian Empire as their cultural heritage and be able to study the sources on which that perspective rests.

From Scholar-Centered Publications to Reader-Centered Infrastructure

Perhaps the most important point of continuity—and the greatest reason why publication in classics has adapted so little to the digital world—appears before we even begin reading the publications themselves. An informal survey of 41 e-classics publications available online from Johns Hopkins University Press reveals that 40 (97.5 percent)⁸ are products of a single author. The only exception was an archaeological publication in *Hesperia*, the journal of the American School at Athens. While expanding this survey would provide greater statistical certainty, the conclusion would be the same: classicists in 2008 devote most of their energies to individual expressions of particular arguments.

Single-author publications will remain important, but even they can adapt to the digital. Athenian democracy was a major cultural event in human history, and it deserves careful study. So much scholarship has accumulated around this topic that recent professional

⁸ This informal survey examined the articles in sample issues that Johns Hopkins made publicly available for marketing purposes. Where there was not a public issue, the most recent online issue was examined. Seven single-author articles in http://muse.jhu.edu/demo/american_journal_of_philology/: 126(1) 2005; five single-author articles in http://muse.jhu.edu/demo/arethusa/: 2005: 38(1); four single-author articles in http://muse.jhu.edu/demo/classical_world/: 2005: 99(1); http://muse.jhu.edu/demo/helios/: 2007: 34(1); nine singleauthor articles in http://muse.jhu.edu/journals/journal_of_late_antiquity/ toc/current.html: 2008: 1(1); two single-author articles in http://muse.jhu. edu/journals/mouseion_journal_of_the_classical_association_of_canada/ toc/mou.7.1.html: 2007:7(1); 10 single-author papers in http://muse.jhu.edu/ demo/transactions_of_the_american_philological_association/: 2005;135(1); and three single-author papers in http://muse.jhu.edu/demo/hesperia/2005: 71(1). By contrast, there was only a single multiauthored paper in this group: J. C. Kraft, G. Rapp, J. Gifford, and S. Aschenbrenner. 2005. Coastal Change and Archaeological Settings in Eli. Hesperia 74: 1–39.

publications cite other secondary sources and often do not cite the primary sources on which our ideas ultimately reside. These publications assume that their readers will either take their conclusions at face value or will have access to extensive research libraries that contain the specialist journals and monographs cited. The authors, their reviewers, and their publishers collectively decided that the benefits of citing primary sources were not worth the cost. General readers would not have access to the primary sources. If they did, they probably would not make the effort to pull them from the shelf. And if they did pull them from the shelf and were able to understand the canonical citation schemes that describe the location of a passage in a text, they would probably not understand what they were looking at. Finally, even if the publishers distributed digital copies of the work on Athenian democracy, the publisher's subscription model would ensure that those publications would reach only those with access to the academic research libraries: many publishers specify that libraries not provide remote access to university alumni and scholars from other, less wealthy institutions.

The top two sites that Google retrieved for "Athenian democracy" in August 2008 were the article in Wikipedia⁹ and "Athenian Democracy: A Brief Overview," ¹⁰ from Demos: Classical Athenian Democracy, a book-length and book-like electronic publication on Athenian democracy, largely written by Christopher Blackwell, a classics professor at Furman University, but including labeled publications from other authors as well. While source files are TEI-compliant XML, the form of Demos is entirely traditional: it consists of expository prose and can be downloaded as HTML and PDF.

Two features distinguish the content of Demos from that of its print counterparts. First, Demos is available as an open access publication hosted by the Stoa Publishing Consortium, founded by Ross Scaife in 1997 (and still in operation after its founder's untimely death in March 2008). Second, Demos was composed from the start to exploit the fact that most of the sources about Athenian democracy are freely accessible online as part of the Perseus Digital Library. Demos thus systematically provides links to the primary sources on which its statements about Athenian democracy are based. Demos also includes information about the cultural context and biases of the various Greek sources so that readers will have the background with which to begin critically evaluating the sources on their own. Demos provides a tangible example of how scholarship can substantively exploit the possibilities of the digital medium.

The juxtaposition of Wikipedia and Demos points to one possible way forward in scholarship. We need to combine the immense cultural energy in community-driven projects such as Wikipedia

⁹ http://en.wikipedia.org/wiki/Athenian_democracy.

¹⁰ http://www.stoa.org/projects/demos/home.

with the intellectual transparency for which Demos strives. ¹¹ While we will need to develop new ways to evaluate scholarly contributions, classicists at least should have little trouble looking beyond the single-author monograph publication model that now dominates in much of the humanities. ¹² Many of us in the field remember when the production of critical editions, scholarly commentaries, and other largely infrastructural projects was still the most prestigious form of publication.

The grand challenges of twenty-first century scholarship reimagine in a digital world the infrastructure that had taken shape to serve the practices of print culture. As early as 1465, Furst and Schoeffer printed Cicero's De Officiis and Paradoxa in Maintz. 13 After 500 years of continuous scholarly development, print infrastructure for classics had reached a considerable level of maturity. The form of our commentaries, critical editions, lexica, encyclopedias, atlases, and other scholarly tools remained unchanged throughout the twentieth century. Even after the TEI had published conventional methods with which to create genuinely digital editions and geographic information systems had begun to revolutionize the ways in which we visualize space, classicists published print editions and maps. And while some of us were eager to exploit such new methodologies at an early stage, few of us anticipated the immense impact and raw utility that projects such as Wikipedia would exert. The assumptions of print publication had so shaped our thinking that we could not believe that such a radically new form of intellectual production would succeed.

We now face the challenge of rebuilding our infrastructure in a digital form. Much of the intellectual capital that we accumulated in the twentieth century is inaccessible, either because its print format does not lend itself to conversion into a machine-actionable form or because commercial entities own the rights and the content is not available under the open-licensing regimes necessary for eScience in general and ePhilology in particular. 14 Even if we care only about

¹¹ Much research has explored how both the Wikipedia model and the data produced by Wikipedia might be useful to the scholarly community. For example see, D. Milne, et al. 2007. A Knowledge-Based Search Engine Powered by Wikipedia. *Proceedings of CIKM 2007*, 445–454; and R. Rosenzweig. 2006. Can History be Open Source?: Wikipedia and the Future of the Past. *Journal of American History* 93(1): 117–146.

¹² Indeed there have been a number of recent challenges to the single-author monograph model as well as a call to reshape the entire structure of traditional scholarly communication, for example, see C. Bazerman, et al. 2008. Open Access Book Publishing in Writing Studies: A Case Study. First Monday 13(1). Available at http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2088/1920. See also H. Van de Sompel and C. Lagoze. 2007. Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication. CTWatch Quarterly 3(3). Available at http://www.ctwatch.org/quarterly/articles/2007/08/interoperability-for-the-discovery-use-and-re-use-of-units-of-scholarly-communication/.

¹³ Sandys, J. E. 1908. *A History of Classical Scholarship*, vol. II. Cambridge University Press, 102.

¹⁴ For further discussion of this issue, see W. Arms and R. Larsen. 2007. *The Future of Scholarly Communication: Building the Infrastructure for Cyberscholarship.* Report on a NSF-JISC Workshop, April 17–19 2007. Available at http://www.sis.pitt.edu/~repwkshop/SIS-NSFReport2.pdf.

our own research, we need content that can be freely analyzed, visualized, and repurposed. And if we want the ancient world to play the most vigorous possible role in the intellectual life of humanity, we want all the results of our work to be physically and intellectually accessible to the widest-possible audience.

We need to build an infrastructure that provides at least three kinds of access:

- Access to digital representations of the human record: This implies providing the best-possible digital representations of our primary data to as many people at as many points on the globe as possible. At this level, we may be delivering a page image from an untranslated Greek text or images associated with some physical location. The term *digital surrogate* is misleading because digital representation such as very high-resolution multispectral scans of manuscripts will often provide more information than would simple access to the physical object.
- Access to labeled information about the human record: We should be able to ask for information that is explicitly stated about any named entity: places (e.g., Salamis in Cyprus versus Salamis near Athens); people (Alexander the Great versus the Alexander King of Macedon, who collaborates with Persia in Herodotus); canonical texts citations (e.g., Greek editions, modern language translations, or commentaries that correspond to lines 11–21 of Book I of Homer's *Odyssey*); linguistic phenomena (e.g., the Greek accusative absolute). This level of access essentially (and dramatically) extends the coverage and precision of existing library catalogs, including domain-specific content.
- Access to automatically generated knowledge: We can use machine-readable encyclopedias with articles about multiple figures with the same name (e.g., different people named Alexander or different places named Alexandria) to analyze the content of these articles for clues with which to determine which of these Alexanders or Alexandrias particular passages in classical texts probably denote. We can use machine-readable dictionaries and modern language translations aligned to Greek and Latin source texts to determine the meaning of a particular word in an untranslated passage (e.g., does Latin orationes correspond to English "prayers," "speeches," or something else in a given passage?). We can use Treebanks (databases that track the syntactic relations of words in a sentence: e.g., word *X* is the main verb, with word Y as its subject and word Z as its object) to train parsers that can then begin decoding the syntactic structure of sentences for which no parses exist. We can use models of a user's educational background (e.g., the vocabulary of every Greek text and the textbooks on ancient Greek history they have studied in their coursework) to predict new words and concepts in a given passage and then to rank these new words and concepts by importance according to various criteria.

Our ultimate goal must be to make the full record of humanity accessible to every human being, regardless of linguistic and cultural background. In this, we expand upon the recurrent and obviously impractical idea of capturing the sum of human knowledge. Similarly chimerical impulses surely were at work in the Aristotelian school of fourth-century BCE Athens, the great library of Alexandria in the third century, the entrepreneurial printers of Europe in the late fifteenth century, and German classicists of the nineteenth century, just as similar dreams move projects such as the Open Content Alliance (OCA) and Google Books in our time. The impracticality of these impulses served the very practical purpose of helping each of these projects envision a radically different world and leave the world different, indeed better, than they found it.15

The universal library represents an unattainable point of reference: it is like a star toward which we navigate. If we face in this direction, we can flesh out the twists and turns of navigable paths toward distant but attainable goals. For our group, the goal is to make the core information about the classical world accessible to speakers of every major European language and of Chinese and Arabic. The European Union has a fundamental mission to serve its own language communities and has made an ongoing investment in multilingual technologies. The United States Government, by contrast, identified Arabic and Chinese as strategic languages. Corporations such as Google, Yahoo, and Microsoft serve global audiences and have major needs for multilingual systems. Classicists can organize their labor to build upon these larger infrastructural efforts.

There are three major strategies to make a growing core of information about the Greco-Roman world accessible to audiences in a range of languages and cultures.

• Domain optimization for machine translation: General systems for machine translation, translation support, cross-language information retrieval, and other multilingual services attempt to do a reasonable job on any category of input, but in so doing, they cannot make simplifying assumptions about the text on which they are working. In effect, we create language models for representative corpora about Greek and Latin. Such language models would reflect the fact that a term such as *case* probably describes a linguistic category (e.g., accusative or dative case) in a grammatical text but not a display cabinet in a museum catalog. A preprocessor could label most likely translations for those terms whose meanings diverge most in a given text from more-general language models. Such an approach requires training data for each source

¹⁵ Much has been written comparing the different models of the OCA and Google Books. See R. K. Johnson. 2007. In Google's Broad Wake: Taking Responsibility for Shaping the Global Digital Library. *ARL: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC*, 1–17; and O. Y. Rieger. 2008. *Preservation in the Age of Large-Scale Digitization*. Washington, D. C.: Council on Library and Information Resources. Available at http://www.clir.org/pubs/reports/pub141/pub141.pdf.

¹⁶ For an example of this work in terms of digital libraries, see M. Agosti, et al. 2007. Roadmap for Multilingual Information Access in the European Library. *Proceedings of ECDL 2007*, 136–147.

- language (in classics, English, French, German, and Italian as a start). Such training sets may require substantial labor to establish but they can be applied to open-ended bodies of semantically coherent text.
- Maximizing the amount of basic data stored in ontologies and other abstract formats: Ontologies can rapidly become complex and idiosyncratic, but if we concentrate on basic propositional statements from mature conceptual reference models (e.g., TEI P5,17 CIDOC CRM,18 FRBRoo19), we can create knowledge bases that are much easier to convert into multiple languages than is full text. The ontological categories should allow systems to apply the classics language models even more effectively than in more general text (i.e., systems will have much better data with which to determine whether they are viewing museum catalog entries or a grammatical database when they confront terms such as *case*).
- **Exploiting detailed linguistic annotations on canonical texts:** Perseus has already published the first 50,000 words of a Latin Treebank, representing the syntax of each sentence as a tree structure and thus addressing one major category of ambiguity that causes problems in machine translation.²⁰ Work continues on the Latin Treebank, and Perseus has just received funding to begin work on a million-word Treebank for classical Greek. Other forms of annotation allow us to resolve additional classes of ambiguity (e.g., a co-reference annotation would allow us to indicate that a pronoun such as *hic* refers to Cicero rather than Caesar). Digital editions may devote more energy to linguistic annotations of this kind than to the traditional revision of textual readings in frequently edited texts. We should design these annotations to facilitate accurate translation into multiple languages. The annotations being keyed to Greek and Latin are, in fact, another form of propositional knowledge and should be useful to anyone reading Greek and Latin, whether they are native speakers of Arabic and Chinese or of English and German.

In producing a digital infrastructure for their field, classicists find themselves engaged again in the most established scholarly practices of their field: the production of editions, lexica, commentaries, encyclopedias, grammars, and other scholarly tools. In the digital world, however, these tools are no longer static objects but dynamic systems that can interact with each other and with their human readers. These books begin to answer Plato's criticism that writing could not answer the questions posed by its readers. Classicists are now in a position to begin new research projects that were not feasible in print culture. Even more important, classicists can now expand the role that their field plays, not only in Europe and North

¹⁷ http://www.tei-c.org/Guidelines/P5/.

¹⁸ http://cidoc.ics.forth.gr/.

¹⁹ http://cidoc.ics.forth.gr/docs/frbr_oo/frbr_docs/FRBR_oo_V0.9.pdf.

²⁰ http://nlp.perseus.tufts.edu/syntax/treebank/.

America but also in intellectual communities with ancient classical traditions (such as the Islamic world) and in which Greco-Roman culture can figure with more prominence than was ever feasible before (such as China and India). Classicists—and all humanists—have an opportunity to develop a new, global intellectual culture that transcends the boundaries of the past.