Classics in the Million Book Library

Christopher Blackwell, Furman University
Gregory Crane, Tufts University
Helma Dik, University of Chicago
Catherine Mardikes, University of Chicago Library
Charlotte Roueche, King's College London
Jeff Rydberg-Cox, University of Missouri
Ross Scaife, University of Kentucky
Neel Smith, College of the Holy Cross

University of Chicago, November 7, 2006[1]

Background

Classicists face a unique opportunity. Google, Microsoft, the Open Content Alliance and other emerging projects (such as the European i2010 initiative) have begun to create very large collections ultimately designed to exceed in size the largest academic print libraries on earth. Classics stands to gain more than many disciplines. Anything published in the United States many useful editions, reference works and publications are in the public domain and thus are among the first texts to be included in these projects.[2] If large entities such as Google and Microsoft are able to provide access to materials protected by copyright, whether by application of fair use or by licensing with publishers, the potential value of these collections increases dramatically.

Classicists have already spent a generation laying the foundations for a digital infrastructure within which to explore the ancient world.[3] While the potential for

[1] This meeting was convened on November 7, 2006, at the University of Chicago as a part of a Mellon funded project, centered at Tufts University, to explore the question, "What do you do with a million books?" A call for papers due December 15 is available at http://www.stoa.org/?p=488. A final workshop will take place at Tufts University, May 22-24, 2007. This position paper is intended as a starting point for public discussion that will inform discussion and May. This discussion will be as open as possible and is designed to attract multiple voices and perspectives.

[2] In the United States, anything published in or before 1922 is now in the public domain and likely to find its way in one of the mass digitization projects. Many works published after 1922 are also in the public domain in the United States, whether because their original copyright was not officially renewed or because a US rights holder did not in some other way follow the requirements of receiving copyright protection (e.g., anything published up through 1977 without copyright notice in the US). For a summary of what is in the public domain in the US, see http://www.copyright.cornell.edu/training/Hirtle_Public_Domain.htm.

[3] The Thesaurus Linguae Graecae, founded in 1972, is the longest project in continuous operation in classics. Sites such as http://www.stoa.org/ and http://www.digitalclassicist.org/ provide a forum that mentions many new projects. Maria Pantelia's web page "Electronic resources for classicists: the second generation,"

research is immense, Classics thus is well poised to make these materials intellectually accessible and coherent to vast new audiences. These vast new audiences include not only the general public but also scholars without access to major research universities – one member of the committee cited a September meeting with classicists from the former eastern bloc who depended almost entirely on freely available scholarly resources on the internet.

Digital publication has already allowed classicists to reach beyond the traditional canon of Greek and Latin materials. Documentary materials such as papyri and inscriptions, previously published, if at all, in expensive specialist publications, are now openly accessible on-line. Later Greek and Latin texts, previously confined to special collections and rare book libraries, now circulate within increasing ease in a digital environment, opening up fields to a vastly wider circle of scholars.[4]

At the same time, in an increasingly digital world, we need scholarly tools created in and designed for the emerging digital world. These tools must, like their print counterparts, serve human readers and include articles, monographs, scholarly editions, introductory materials and annotations, encyclopedia entries, and similar analogues to print publication. They must also take creative advantage of the new opportunities granted to networked scholars working with rich digital resources.

Immense digital libraries based on open access and aimed at massive audiences put scholars under an obligation to avoid a new access divide opening up between ourselves and the wider community that we serve. Peer review, distribution, production and structure of new open source editions are the object of vigorous, on-going international discussion. For example, the UK Arts and Humanities Research Council funded a recent workshop on Open Content Scholarly editions that brought together scholars from Europe and the United States.[5] This workshop brought together many projects already engaged in creating open source scholarly editions. One member of this committee is an author of a separate report to advance peer review, funded by the AHRC, which is due to be released shortly. A follow up meeting will take place in 2007 to formalize positions articulated in this workshop and at other international venues.

provides a sense of the activity in this field but (as of November 13, 2006) had not been updated since September 9, 2004: http://www.tlg.uci.edu/index/resources.html. It is unclear whether any one individually can manually track digital work in classics at this point.

[4] The Camena/Termini Thesaurus Eruditionis provides an example of purposeful collection development that shrewdly integrates page images and XML transcriptions, source texts and reference works: http://www.uni-mannheim.de/mateo/camenahtdocs/camenaref_e.html; http://www.dlib.org/dlib/march06/schibel/03schibel.html. There are dozens and probably hundreds of scholarly projects underway that are placing previously inaccessible Greek and Latin materials in wider circulation. Dana Sutton's *Analytical Bibliography of Neo-Latin Titles*, current through November 10, 2006, lists 19,900 records.

[5] http://digitalclassicist.xwiki.com/xwiki/bin/view/Main/osce.

The expectations for digital editions should be higher than for their predecessors: we expect dynamic textual notes (compare witness A vs. B), links to high resolution images of the manuscript, papyrus, inscription or other source,[6] and potentially even new forms of annotation (e.g., syntactic markup as a component of a standard edition). Our job is to make these huge new collections a foundation for this next generation of more expressive and sophisticated editions.

New digital tools should go beyond their print counterparts in at least three ways.

First, reference materials and scholarly editions should provide a knowledge base to support advanced services: collation and visualization of multiple editions of a single text, machine translation, automatic identification and alignment of translations with their source texts, machine translation, document summarization, morphological analysis, named entity identification, information extraction, text mining and similar services.

Second, reference materials and scholarly editions should be updatable in a continuous, documented, versioned fashion. Each change should have an identifiable source. Citations must be able to reconstruct the knowledge source as it existed as the time when the citation was created and then to trace even complex subsequent changes executed over long periods of time.

Third, human and machine decisions should reinforce each other. Just as OCR should provide a first draft that editors may correct, named entity identification, syntactic analysis, morphological analysis and similar processes should provide useful initial results that human readers can correct and augment. The result should be an on-going interaction, where machine analysis prepares scalable results while human feedback not only addresses specific decisions but also improves the underlying statistical models on which subsequent automatic analysis depends. We need a new generation of tools that can facilitate this intensive interaction between automated processes and expert editorial control.

Recommendations:

1) We should do our best to build on what Google, OCA and other projects are doing, augmenting and enhancing it for the uses of classicists. In our view, this has implications for scholarship and for those broader communities which scholarship serves. Google, OCA etc. are creating collections far larger than classicists could ever construct on our own. Centuries of classical scholarship and a generation of digital classics have put us in a position where we can add value to these raw materials, making them intellectually accessible to audiences at any point on the globe.

[6] http://www.papyrology.ox.ac.uk/POxy/monster/demo/Page1.html; http://www.teuchos.uni-hamburg.de/.

Enhancements may take several steps. First, we must establish a service with materials on which we can freely experiment (these include many of the image books entering the Open Content Alliance). Second, academic libraries such as Michigan can apply these techniques directly to the sources files which Google has digitized from their collections. Third, these services, once published, will, we hope, become standard within large commercial libraries.

Enhancements applied to large collections would include (in roughly sequential order):

a. Enhanced OCR for Greek and Latin. OCR provides serviceable text with which Google can search English. OCR tuned for modern languages often introduces errors of its own into Latin (e.g., “operis” → “opens”, “t-u-m” → “turn”). At the moment, the OCR software used by Google (and OCA) cannot generate any text from classical Greek. We need better OCR for Latin, while any OCR for Greek would be an improvement. Perseus can now generate OCR of Greek that is up to 99.94% accurate, while its morphological analyzer and word lists for Greek and Latin can enhance error correction.[7] Google has chosen an open source OCR package,[8] presumably to facilitate just such contributions from the many communities, most small in themselves, who comprise the long tail of academic life.

b. The application of established techniques for language matching to identify Greek and Latin within the larger collection. We need to provide the tools whereby documents and even short passages in Greek and Latin within these vast collections can be identified, creating scalable collections which grow in size as the overall collections develop over time. We would thus find not only smaller texts embedded in larger anthologies but also the entries in commentaries and even quotations of Greek and Latin passages. The result would be an automatically collated library of full editions and the extracted testimonia (which scholars have painstakingly constructed by hand). Thus, passing quotations of Greek and Latin in later texts will often be matched to their uncited (and untranslated) sources.

c. Text alignment. Some pieces of Greek and Latin will be unique, but many will be editions or excerpts of the same works. We want to be able to recognize multiple editions of the same work, distinguishing legitimate variants from errors, identifying the citation schemes and main document structures involved. This stage automatically compares and summarizes

[7] On the particular problems of early modern books, see http://daedalus.umkc.edu/incunables/index.html; http://www.dlib.org/dlib/march06/choudhury/03choudhury.html.

[8] http://www.iupr.org/projects.

the differences between multiple editions of the same work within the larger collection.

d. Named entity identification. We should be able to recognize names of people, places, organizations in Greek, Latin and the languages of scholarly publication. We should also be able to distinguish one Caesar from another. In the short run, we may have to build such services ourselves. In the long run, we may be able to create knowledge bases that Google etc. can upload, analyze and use to drive more sophisticated services.

e. Linguistic analysis: This may include language specific morphological and syntactic analysis modules designed by classicists and used by Google etc. as components in more general systems.

f. Other forms of text mining. These include classification based textual analysis (e.g., identify features distinguishing male and female speakers in a given corpus) and exploratory clustering (organize a set of documents into groups with shared vocabulary and then analyze the similarities that underly the emergent organization).

g. Reading support. This includes automated processes such as the morphological look-up, dictionary links, automatic keyword linking and links to explanatory materials long established in Perseus, as well as emerging technologies such as machine translation and personalization aimed at providing background suited to particular people at particular times. At the same time, we need mechanisms whereby readers can refine and/or augment machine generated support. These user contributions already include in the Perseus Digital Library the ability to vote for a particular morphological analysis but should cover every machine generated recommendation and should include a wide range of annotations, including translation/glossing of difficult phrases, background on particular passages etc.

h. Authorial support. This includes tools whereby individuals can refine, augment or add materials within the larger collections. Such materials may include new editions, translations,[9] annotations, and extended analyses as well as "micro-publications" (e.g., selecting the correct morphological analysis for an ambiguous word in a given context). Many systems are emerging to support such collaborative work, supporting both contributions with a single authorial voice (e.g., Blogs) and community driven documents with many authors (e.g., open systems such as Wikipedia.org and more managed environments such as Planetmath.org). We need to make sure that these more general systems are able to manage

[9] http://www.stoa.org/sol/.

the nuances that we need in classics: e.g., support for common ways to describe the canonical text citations (e.g., mapping Thuc. 1.38 to Thucydides, History of the Peloponnesian War, book 1, chapter 38),[10] the names of geographic names (e.g., how do we distinguish one Alexandria from the next?),[11] etc.

2) Core open source content: While we may rely on Google etc. to digitize the vast majority of content, we must as a profession take responsibility for creating and maintaining a rich core of reliable editions and reference works described above that can be uploaded into any digital library, and which individuals and groups can use to provide a starting point for the new, open source knowledge base on which scholarship will depend. In this case, the Open Content Alliance (http://www.opencontentalliance.org/), with its commitment to the free distribution of ideas, provides a natural collaborator. Besides its large scale scanning services, it is preparing a cost-recovery mechanism whereby users can select particular books for inclusion within the OCA.

3) Open source services: e.g., GATE (Generalized Architecture for Text Engineering), Zotero, Canonical Text Services, Morpheus, the University of Chicago's PhiloLogic, TextGrid, EpiDoc (discipline-specific XML for epigraphy), Duke Databank of Documentary Papyri, components of the Perseus Digital Library, interoperable geo-referenced data via protocols disseminated by Pleiades.

4) Aggressive outreach. We need to continue educating our colleagues, developing collaborations, broadening support. We need to support international, decentralized, collaborative development in ways that enfranchise the greatest possible number of dispersed scholarly participants.

The potential benefits of these vast new collections for intellectual life, within the academy and beyond, are immense but we will only realize them fully insofar as we ourselves take an active role in shaping this new future, both by articulating our values and by implementing our beliefs.

[10] http://chs75.harvard.edu/projects/diginc/techpub/cts.

[11] http://icon.stoa.org/trac/pleiades/.