Comments on the "APA Task For on Electronic Publications: Issues and Recommendations for Discussion (draft of October 20, 2006)[1]

Gregory Crane
Professor of Classics
Winnick Family Chair of Technology and Entrepreneurship
Editor in Chief, Perseus Project

December 20, 2006

A Joint Task Force on Electronic Publishing of the American Philological Association (APA) and the Archaeological Institute of America (AIA), Donald J. Mastronarde, Chair (September 2006)[2] has produced recommendations on e-publishing. The topic is important and it is good to see that the American Philological Association is contributing to the on-going discussion about the digital future of the field.

The charge of the committee is too narrow. E-publishing is only one component – and a subordinate component – within a larger, evolving ecology for intellectual life. Debate has moved from electronic publication to broader issues such as e-science, cyberinfrastructure, and grid technologies, which roughly correspond to practice, the architecture, and concrete technologies.[3] The center of gravity for intellectual life in academia and society as a whole has already shifted decisively to a digital environment. New forms of producing, disseminating, and preserving knowledge such as blogs, wikis, institutional repositories, encoding standards, ontologies, etc. are providing a foundation for post-incunabular digital libraries that do more than shuffle static PDF imitations of print. Other humanists have suggested that we need to explore the more general topic of e-research.[4] The American Philological Association might consider e-philology and how a broad *scientia totius antiquitatis* can evolve in a rapidly and radically changing world.[5] Before making recommendations about electronic publication, the field needs to debate how evolving audiences for, contributors to, and forms of publication relate to our strategic goals.

---

[1] http://socrates.berkeley.edu/~pinax/taskforce/APATaskForceDiscussionDraft3.pdf.
[2] These are available http://www.apaclassics.org/Publications/e-publishing.html.
[3] Lynch, C (2006) Research libraries engage the digital world: A US-UK comparative examination of recent history and future prospects. *Ariadne,* 46 http://www.ariadne.ac.uk/issue46/lynch/.
[4] http://lists.arts.usyd.edu.au/pipermail/e-humanities/2005-September/000152.html.
[5] For an extended discussion of e-philology by the author, see http://repository01.lib.tufts.edu:8080/fedora/get/tufts:facpubs.gcrane-2006.00003/bdef:TuftsPDF/getPDF, a preprint of "ePhilology: when books talk to their readers," forthcoming in the *Blackwells Companion to Digital Literary Studies*, Ray Siemens and Susan Schreibman, eds (forthcoming 2007). For a related discussion of post-incunabular digital libraries with classics as an example, see "Beyond Digital Incunabula," http://repository01.lib.tufts.edu:8080/fedora/get/tufts:facpubs.gcrane-2006.00002/bdef:TuftsPDF/getPDF.

The ACLS/Mellon funded Commission on Cyberinfrastructure for the Humanities and the Social Sciences has been circulating draft reports since 2005 and on December 16, 2006, issued its final report.[6] This document now provides the framework for discussion of a digital future in classics and in any humanities discipline. Until we have an idea of how we fit into this larger scheme, we are not in a position to make informed plans or to propose funded projects of compelling interest.

The natural starting point for the American Philological Association would be the five goals by which the APA defines itself to the public:[7]

- Reassert the importance of primary and secondary school teaching and provide more support for improved pedagogy at all levels of teaching.
- Improve working conditions and scholarly opportunities for university and college teachers.
- Increase communication with audiences beyond its membership.
- Make sure the Association's research program is appropriate to the needs of the profession.
- Coordinate and systematize data collection in order to provide an accessible and reliable information base to support other Association goals.

The cyberinfrastructure that evolves will constrain the extent to which we can exploit the three initial goals outlined above.

First, if we want to support teaching, we need a cyberinfrastructure to support those working with Greek and Latin source texts, in the original and in modern language translations. The challenges of historical languages differ in many ways from those of modern languages. Classicists must take the lead to make sure that their needs are met, but in so doing they must actively invite collaboration with colleagues in classical Arabic, classical Chinese, Sanskrit, Old Norse, Middle High German, Akkadian, Sumerian, and every historical language.

Second, one way to improve working conditions and research opportunities for university and college teachers is to support, in every possible way and with as much energy as we can muster, the creation of massive digital libraries based on open access such as those now being built by Google, Microsoft, the Open Content Alliance and others. Even if only partially realized, these efforts will expand the intellectual reach of all college and university teachers. If these efforts come close to their original goals, we will find on-line and freely accessible a larger and far more useful research library than any institution of higher learning has ever created. Classicists stand to gain more than any other discipline, for the field is often strongest at liberal arts colleges which have never had

[6] http://www.acls.org/cyberinfrastructure/cyber.htm.
[7] http://www.apaclassics.org/about.html.

access to first class research environments. Nor is open access alone always enough. After sustained requests from an increasing set of researchers, we at Perseus decided to make all the content that we could available under a Creative Commons attribution/sharealike/non-commercial license. Researchers want to apply their own analytical tools to the full source texts and to create derivative works. All of the rising scholars mentioned later in this piece have vigorously argued for an open sources rights regime, where all have equal access *sine ira et studio*. The growing desire for such resources extends beyond the United States: UK Methods Network funded an initial conference on Open Source Scholarly Editions for classics in September 2006.[8]

The third goal is, arguably, the most important. If we want to expand the role of classics in teaching and to improve the lives of classicists in higher education, we must expand our role in the intellectual life of academia and of society as a whole. By making our content intellectually more accessible, we can enhance its value for students, for researchers beyond classics, and for society as a whole. Certainly this audience exists: where the membership of the APA stands at c. 3,100, in our last survey (April 2005), we found 400,000 unique users of the Perseus Digital Library. Of these, c. 30,000 were working directly with source texts in Greek and Latin. This sample reflects only a subset of our potential current audience. If we are able to foster an infrastructure that carries our content out across the world and makes it more useful to expert and novice alike than was ever possible in a purely print world, then we will have the tools to build a field that is more vital and contributes more to the intellectual life of humanity than at any time in the past.

The fourth goal depends upon the "needs of the profession." At some level, our goals remain unchanged regardless of the technological infrastructure, but our real world needs to reflect opportunities and challenges that change along with technological infrastructure. We also need to identify the different needs within the field. To what extent do we maintain the status quo with marginal changes? If the practices of intellectual life and the methods of creating and sharing ideas are changing, what are the implications? How do we create an environment that not only serves current members of the profession but that attracts the best emerging talent to the field?

The following section addresses, in part, the fifth goal. The conversation about digital classics needs to include not only experts in other areas but the generation of classicists who have emerged over the past fifteen years and who represent a first wave of scholars who grew up in a partially digital world.

Expanding the conversation

To advance this discussion, the APA should engage experts within classics and beyond. Within this landscape the million book question may be the most important for classicists: What happens as very large collections take shape? What are the prospects for scholarship? What technologies are emerging that may enhance the intellectual

[8] http://www.stoa.org/?p=484.

access to these collections? What are the tactical implications for a field such as classics? The Mellon Foundation has funded a study on this topic that will conclude in the spring of 2006.[9] A group of classicists met in Chicago in November 2006 to publish an initial report suggesting the implications for the field.[10]

Plans for cyberinfrastructure or publishing must consider the technologies with which we will locate, browse, analyze and visualize digital publications. Rising computer scientists with classics backgrounds such as David Smith at Johns Hopkins and David Mimno at UMass Amherst are in a position to help us understand the state of the art for machine translation and text mining and the role that these technologies can play in classics. Helma Dik of the University of Chicago helped organize an important conference on Digital Humanities and Computer Science that established new connections on which classicists can build.

Classics needs to consider seriously the on-going major digitization projects. Relegating them to the appendix under the title "old books" because "it is not clear how such projects will work out" is problematic. Given the scarcity of resources that this report stresses, how can classicists undertake projects that may be rendered irrelevant? It is hard to see how Google/Yahoo can be given five lines in an appendix and the idea of digitizing twenty-year old microfiches of "old books" get the better part of a page in the main report. This APA draft report effectively assumes that these huge projects will either fail or be irrelevant. Subsequent plans must consider the implications for classics if these projects are partially or wholly successful.

The **Google Library Project**: The University of Michigan has provided the core collection on which Google is building – it opened all of its materials, not just a subset, to Google from the start. Michigan retains rights to include materials that Google digitizes in its own digital library services and to share that content with other academic libraries. Michigan also included a clause in its Google agreement ensuring that all Michigan materials on-line be available without cost to the end user.[11] It may not be, as the October 20 draft suggests, "clear how such projects will work out," but even partial success will fundamentally change our world.

* If Google does succeed in scanning all of Michigan's collection and searches simply return bibliographic records for every journal article in the Michigan library, what will be the implications for classical scholarship? What will happen as increasingly sophisticated text mining and visualization software classifies and discovers patterns in the most important publications of classical scholarship?

*Recent experiments with Greek OCR at the Perseus Project have produced astonishingly good results: we have found that modest training of commercial OCR engines produces transcriptions of Greek character data that are 99.72% accurate on a clear 20th century text. Using morphological analysis software as the foundation for automatic error

[9] http://www.stoa.org/?p=488.
[10] http://www.stoa.org/?page_id=516.
[11] http://www.lib.umich.edu/mdp/umgooglecooperativeagreement.html.

correction, accuracy rises to 99.94% -- barely below the 99.95% standard in professional data entry contracts. What will be the implications of being able to search not only the full text but variants and multiple editions of Greek and Latin texts? What will be the implications if we can find quotations of Greek and Latin in reference works and secondary sources and then link these to the original sources?

Classics is fortunate in that Derek Collins, of the University of Michigan, spent an extensive amount of time on the search committee for the University Librarian. The Google collaboration is the signature project at Michigan. Derek Collins is thus in a unique position to help classics learn what Google and Michigan both are doing.

The **Open Content Alliance**: The Google Book project takes open access as a core feature of its business plan – in effect, Google has decided that it will make more money reaching a global audience than by selling subscriptions to a smaller community. But if Google has adopted an open access policy, its texts are not open source. Google is not worried about scholars – the Google libraries can freely make their scanned images available in a non-commercial form via their own searching and digital library systems. But Google does not want major competitors such as Microsoft and Yahoo to benefit from its substantial investment in digitization.

The Open Content Alliance, founded by Brewster Kahle and centered around the Internet Archive, has created an efficient work flow to create high quality, open source image books that anyone can download and freely repurpose. Scanning centers are in operation at Toronto, the University of California (two separate centers) and other locations. With scanning costs at $.10/page and set-up costs of $5 a book, the Open Content Alliance can digitize a book and make it available to the world for less money than the print purchase price to which classicists are accustomed. Individual scholars could conceivably select and pay for books to be added. The Perseus Project has already scanned c. 1,000 volumes of Greek and especially Latin editions, commentaries and translations for dissemination as image books. The Johns Hopkins University Libraries has recently begun contributing to this effort.

Classicists should consider whether there are publications that should be available in high quality for download so that they can be analyzed with newer software or simply read at our convenience. New companies such as Lulu.com and Booksurge.com are emerging that can create high-quality bound paper books from these image books. Should we encourage a demand print library that could include and print all public domain and specifically licensed materials that are under copyright?

Contacts include Brewster Kahle, the founder of OCA (Brewster@archive.org), as well as Chet Grycz, Curator of Books at the Internet Achive (grycz@archive.og) and Sayeed Choudhury, associate director at Johns Hopkins (sayeed@jhu.edu).

Those doing pioneering work advancing classics in a digital age must also be drawn into the conversation. While the October 20 draft mentions established projects such as the

TLG, BMCR, and APh, it does not cite, except briefly in an appendix, any of the major efforts of the past fifteen years. Voices that need to be consulted include:

- Helma Dik of the University of Chicago is developing a corpus based grammar of classical Greek. This not only goes beyond traditional grammars by its use of quantifiable data but is designed to be customizable, such that different audiences can dynamically choose to see different configurations of data. She is also leading a collaboration between Perseus and the French ARTFL project at Chicago.[12] This collaboration is significant because (1) it dramatizes the need for open source (not just open access) publications (2) it reveals the need to design documents which can be published in multiple systems.

- Thomas Elliott for years led the development of EpiDoc – a standards based guideline for electronic publication of inscriptions.[13] For years the director of the Ancient World Mapping Center at UNC, he now directs the Pleiades Project. Funded in 2005 by the NEH, Pleiades is a new open source publication environment for Greek and Roman geography.[14]

- Thomas Martin's *Overview of Greek Culture*[15] and Christopher Blackwell's *Demos*[16] were not just on-line publications but were also the first humanities monograph that took seriously the implications of having access in the same digital space to a library of primary sources.

- Bruce Robertson has done foundational work on ontologies for classical and humanistic studies. His Historical Event Markup and Linking Project (http://heml.mta.ca/) creates a framework not only for classicists but applicable to a wide range of languages and disciplines. His work exemplifies how classicists can advance their own field by asking questions that go beyond and create bridges with other disciplines.

- Jeff Rydberg Cox conducted foundational research on language technologies in a joint European Union/National Science Foundation project.[17] Among other things, he produced the working environment on which the new Cambridge Greek Lexicon depends and which will allow classicists to see not only the articles but the slips on which they are based. Jeff thus advanced the form of a core reference work.

- Ross Scaife's contributions with the Stoa.org in general need to be part of the discussion both because of his own work and his contributions to associated projects. The Suda On Line that has translated more than 20,000 entries of the

[12] http://www.lib.uchicago.edu/efts/PERSEUS/.
[13] http://www.stoa.org/projects/epidoc/stable/guidelines/.
[14] http://www.unc.edu/awmc/pleiades.html.
[15] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:1999.04.0009; jointly published in print by Yale University Press in 1996 as *Ancient Greece: from prehistoric to Hellenistic times.*
[16] http://www.stoa.org/projects/demos/home.
[17] http://www.chlt.org/; http://www.dlib.org/dlib/may05/rydberg-cox/05rydberg-cox.html.

Suda was a pioneering example of new collaborative scholarship.

- Mark Schiefsky and his contributions to the Archimedes Project (http://archimedes.fas.harvard.edu/) has for years been developing methods with which to integrate Arabic into classical studies. His recent paper, "New technologies for the study of Euclid's *Elements*,"[18] summarizes some of his key findings and is essential reading for this topic.

- Neel Smith and Christopher Blackwell's *Canonical Text Services* protocols[19] are one of the most advanced instruments available in any discipline to enable grid services: their work has given classics a huge advantage as we carve out a space for ourselves in an emerging cyberinfrastructure. These are foundational tools for the reference matching that classicists need and that no scientists will provide for us.

The above list only mentions some of the classicists from North America who have helped lay the foundations for a revision of the production, dissemination, usage, audience, and long term maintenance of scholarly communication.

Comments on the October 20, 2006 recommendations

Bibliographic accessibility of electronic resources:

Portals are expensive to create and to maintain. How much will this cost? What will be the benefits for what amounts of funding? How will we pay for it? Is this the best way to invest our scarce resources?

How have other similar projects fared? The portal as described seems to resemble well-established projects such as EdSitement.neh.gov and the massively funded National Science Digital Library (nsdl.org). The report cites "merlot.org for one general effort in this direction." A check of the humanities section of Merlot.org[20] returns a promising 2,479 results. Why not begin by leveraging one of these efforts?

What exactly will this portal contain? Drills for high school Latin? Resources about particular authors? What technological methods will the portal employ?

Should the APA instead be helping foster standards-compliant educational resources that can be sustained over time? Institutional repositories such as Fedora and D-Space and on-line learning environments such as Sakai are emerging to meet the need for scalable, sustainable learning materials.

---

[18] http://archimedes.fas.harvard.edu/euclid/euclid_paper.pdf.
[19] http://chs75.harvard.edu/projects/diginc/techpub/cts.
[20] http://www.merlot.org/merlot/materials.htm?keywords=&category=2327.

Recommendation: Supporters of a portal for web resources should present a proposal for public discussion by members of the American Philological Association. The proposal should present costs, benefits and define how this APA sponsored effort builds upon, and relates to, other projects and best practices.

On-line accessibility: free vs. subscription:

Saving money is a good way to go bankrupt. The strategic challenge that classics faces as a field is to maintain and expand its role in the broader intellectual life of society. Our situation is much closer to that of a university competing with other universities than it is to Hollywood or the music industry raising capital from mass market sales. Classicists have a very small set of primary resources, well-defined and easily stored. We should be focusing on how to augment the role that our subject plays, not only in academia but in society as a whole. The APA reports 3,195 members.[21] In our last survey of users (April 2005), 400,000 unique users consulted the Perseus Digital Library, 90% of whom were working with classical materials. Of these, 10% were using dictionaries, morphological lookups and other indications that they were interacting directly with Greek and Latin source texts. Even if we look only to the narrow interests of the self-selected 3,195 members, we should be focusing our attention on the wider audience on whose interest most classicists ultimately depend.

We need to examine the costs of producing and maintaining scholarly content in our field.

Core production of individual scholarly works: Faculty salaries and sabbaticals pay for the production of content in classics. It would be useful to know how much we depend upon fellowships. All of those who pay our salaries and provide us with research support need us to maximize the impact of our work. To what extent do subscription barriers advance that goal?

Preservation and access: Libraries maintain our publications over time. Publishers drop publications from print to suit their economic interests. Institutional repositories such as D-Space and Fedora have emerged to provide the technical foundation whereby libraries can maintain digital content over long periods of time. The Cornell library now maintains the historic e-print archive arXiv.org. Technologies such as LOCKSS ("lots of copies keeps stuff safe") provide geographically distributed copies. The cost of maintaining digital files over time (which includes, of course, migration to new media) is declining. Books may have long shelf lives but the shelves are expensive to maintain and their costs are not decreasing.

Editorial and formatting costs: Honoraria for editors would be useful but are not essential. Copy-editing is useful but needs revision in a digital environment where machines as well as people are our audience: copy-editing could make documents much more useful if it assured that every bibliographic entry was linked to professional library

[21] http://www.acls.org/aphilola.htm sets the figure at 2,775 individual and 320 institutional memberships (December 14, 2006).

cataloguing data, every Alexandria was associated with the right Alexandria as designated in an authority list such as that developed by the Barrington Atlas and every source citation (e.g., "Thuc. 1.86") could be automatically linked to a text (e.g., Thucydides, History of the Peloponnesian War, Book 1, Chapter 86). On the other hand, automated systems can extract most of this information. Does the benefit of manually adding this information – or even checking the results of automated analysis – justify the cost? In any event, these additions are aimed at making documents more useful as parts of modern digital libraries. Libraries may thus be better suited to this task than publishers.

Library acquisition budgets provide the foundational support on which our publications depend. The challenge lies in shifting library investments away from paying publishers to supporting their own faculty publications. The real question is why we need to invest any hard money in this stage and why classicists cannot review and edit each other's publications as a part of their standard work.

New infrastructure: Some projects go beyond the capacity of individuals or small groups. Luckily for humanists, billions of dollars are being invested in an infrastructure which, while designed for scientific publications and datasets, provides a reasonable foundation on which we can build. We need to identify the smallest possible subset of services that humanists in general and classicists in particular must provide. While funders such as the Mellon Foundation and the National Endowment for the Humanities can help spur development, long term sustainability depends upon support from libraries and academic technology groups.

Capital Intensive Content Creation: Scholars created our critical editions by hand over generations. Digitizing core resources in a short period of time was, however, an industrial process that classicists could not accomplish with the labor at their disposal. Classics thus raised money from NEH and other sources to create the TLG. Commercial vendors subsequently entered the market and created a growing set of databases, drawn from public domain data but available only through subscription.

Another funding model emerged at the University of Michigan. The Text Creation Partnership creates consortia of libraries that band together to produce content which they – and not commercial vendors – own. The Early English Books Online Text Creation Partnership[22] has, since 1999, digitized 700 million words of text, a figure that will probably exceed 1 billion words when the project concludes in c. 2009. These files are available today for all scholars at institutions that belong to this consortium. Five years after the project is completed, the files will be handed into the public domain. The TCP thus constitutes a pragmatic compromise between the need for open source data and the need to extort financial support from libraries.

[22] http://www.lib.umich.edu/tcp/eebo/.

3-4. On-line accessibility: inclusions and exclusions and Conferences

These recommendations seem fine as is.

5 Digital Monograph Series

It is unfortunate if OUP-USA has the right of first refusal to putative digital APA monographs.

Whether the APA can directly sponsor a monograph series or not, it could help advance the question of what digital monographs might look like. Are they simply digital versions of print book-length publications? If so, are they just unstructured text in PDF or should they aim at the higher functionality that XML markup brings? Should they include data sets where applicable? We need to prepare for publications that cannot exist in print: e.g., a dynamic apparatus criticus that would allow readers to compare and analyze multiple witnesses and editions. The first treebanks – databases of syntactic data – for Latin have begun to emerge in Europe and the United States.[23] What kinds of publications will we need to flourish in the coming years?

6 Digital Collection of classical research resources

It is not clear how members of the field are supposed to comment on this topic without a list of what these microfiches contain. A link to an on-line version of the 1985 catalogue needs be added before any discussion with members of the APA can begin.

The Open Content Alliance provides an open source scanning project with scanning centers at a growing number of universities. Books in these library collections can be scanned at high quality with costs estimated at $.10 per page and $5 for shelving. Would it be more useful to sponsor clean scans from some or all of these older publications? If fresh scans provide a better source for Optical Character Recognition software, new scans could dramatically increase the realized value of these works.

Increasingly classics books – including editions and commentaries – are turning up in Google Book search. Google has begun allowing users to download PDF versions of their scanned books when these are in the public domain. What are the implications of this real and present project for the potential value of scanned microfiches?

7 Funding sources

Classicists need to develop alliances with other, better-funded disciplines. We need to find opportunities to build on on-going and new project funded by NSF and NIH. The

[23] The Perseus Project has begun some initial work in this area please see, "The Design and Use of a Latin Dependency Treebank." http://geryon.perseus.tufts.edu/data/%5CPapers%20and%20Props%5C110.pdf

most important funding agency for the future of classics may well be DARPA, the Defense Advance Research Projects Administration: emerging language technologies such as information extraction, text mining, machine translation etc. are redefining the work on which linguistic study depends. Classics is a special case of corpus linguistics and must make intelligent use of what the computational linguistics write and build.

The most important sources for funding are probably in Europe. The British Arts and Humanities Research Council, the Deutsche Forschungsgemeinschaft of Germany, the national funding agencies of each European country and of the European Union as a whole should be part of any field wide efforts to establish a new infrastructure for classics.