Building a Digital Library: the Perseus Project as a Case Study in the Humanities

Gregory Crane
Tufts University

Abstract

This paper outlines some of our preliminary findings in the Perseus Project, an on-going digital library on ancient Greek culture that has been under development since 1987.

The Perseus Project is a collaborative effort to which many colleagues from institutions across the country have contributed.[1] Since the earliest planning stages in 1985 and the beginning of serious development in 1987, its goal has been to create a "critical mass" of materials about a single, but complex, domain: ancient Greek culture. Even ten years ago, there was a growing consensus that ever larger bodies of material would be available on-line and that these would change the way in which we frame and pursue questions, but we believed from the start (and experience has reinforced this belief) that it is almost impossible to predict with accuracy what electronic tools will and will not prove valuable: even the simplest translation from print to bits can transform the relationship between the content and the user in ways that no one had anticipated. Services that seemed promising have at times been less useful, while functions which had attracted little or no thought have emerged as dynamic new tools and even catalysts for the transformation of practice (often a far more influential and difficult transformation than the shift from one theory to another).[2] We wanted to provide one well developed example of a compact, but broad, digital library on a particular domain so that others would be able to turn to it as a concrete example.[3]

After initial planning, we chose to concentrate on collecting primary materials and developing an infrastructure on which others could build rather than on creating a unified curriculum or megatext — in effect, we chose to build a small library rather than a large book. While this decision attracted criticism,[4] it has nevertheless begun to bear fruit: many different HTML documents have already begun to embed links to the growing Web version of our database, and we expect this trend to continue, as on-line journals and publications include links to those source materials that we have placed on-line.[5]

Our intention was to create an extensible, platform-independent infrastructure that could grow over time and that would serve as a foundation for the electronic equivalents of articles, books, presentations, assignments and other forms. As we put it at the time, we wanted to produce a compact library that would not only be useful in and of itself but would also serve as a medium within which to explore the possibilities and challenges posed by a digital library.

In 1992, Yale University Press published Perseus 1.0, which consisted of a CD ROM and optional videodisk. Yale will begin in 1996 distributing Perseus 2.0, which contains a narrative overview of Greek culture (in effect, a 300 page book with hundreds of hypertext links to the database),[6] 3.4 million words of Greek source texts (c. two thirds of all literary texts surviving up through 300 BCE) with accompanying translations, a 35,000 entry lexicon, 1,000 digital maps, 500 plans and drawings, essays on various subjects, several thousand short "glossary" entries, and 24,000 color digital images illustrating Greek art and archaeology. Both Perseus 1.0 and 2.0 are distributed using Apple's Hypercard software, but all

[1]Primary funding for the Perseus Project has come from the Annenberg/CPB Projects; major additional support has been provided by Apple Computer, the National Endowment for the Arts, the National Science Foundation, the Fund for the Improvement of Postsecondary Education, the National Endowment for the Humanities, the Packard Humanities Institute, and the Xerox Corporation. Centered at Harvard University from 1987 until 1993, the Project has since moved to Tufts University. Major development work over the years has also taken place at many institutions, including Bates College, Boston University, Bowdoin College, the College of the Holy Cross, St Olaf College, the University of Chicago, and the University of Maryland.

Permission to make digital/hard copies of all or part of this material for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copyright is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires specific permission and/or fee.
DL'96, Bethesda MD USA
© 1996 ACM 0-89791-830-4/96/03..$3.50

[2]This is a major topic of McLuhan, 1964; for examples of this phenomenon from our work in the Perseus Project, see Crane, 1991.

[3]The need for such testbeds seems only to have increased. On the vagaries of the term digital library, see Fox, Aksyn, Furuta and Leggett, 1995.

[4]For an example of this, see Lee T. Pearcy's contribution to Wiltshire, Pearcy, Hamilton, O'Donnell and Eiteljorg, 1992 (with our response: Crane, 1992): both of these documents are available on the internet at the current *Bryn Mawr Classical Review* gopher site: gopher://gopher.lib.Virginia.EDU:70/11/alpha/bmcr.

[5]Perhaps the best example of this would be the Diotima: "Materials for the Study of Women and Gender in the Ancient World" (http://www.uky.edu/ArtsSciences/Classics/gender.html), created by Suzanne Bonefas and Ross Scaife.

[6]This overview was written by Thomas Martin of Holy Cross College and will be published by Yale University Press in print as well as electronic form.

subsequent releases will be platform-independent. Since early 1995, a Perseus Web site has emerged (www.perseus.tufts.edu). While HTML 2.0 (with its various extensions) and Netscape 1.1 have not allowed us to replicate all the functionality embedded in the Hypercard version, Netscape 2.0 promises to eliminate many of these barriers. The major obstacles preventing a fully accessible Web Perseus are legal and financial rather than technical: we do not have rights to make all of our materials freely available over the Web and we need to protect the investment that Yale University Press has made in the distribution and support of the CD ROM Hypercard based Perseus 2.0. Nevertheless, we are in a position to begin studying the same digital library as it appears in two fundamentally different environments.

According to the taxonomy outlined by Nürnberg et al.,[7] we have spent a great deal of time assembling "data" (i.e., texts, pictures) and "metadata" (i.e., indices and other finding aids). At the same time, the data and metadata are important because they allow us to see how "new digital processes" might differ from "physical library processes." Our initial hypothesis was that a hypermedia environment would encourage people to make connections that the logistical barriers of traditional printed tools had discouraged. Our intention was thus from the start twofold. On the one hand, we wanted to help our traditional audience — students and faculty studying the literature, history, archaeology, art, philosophy, science and other aspects of ancient Greece — pursue their traditional goals more effectively: indeed, unless we could solve existing problems and thus attract people to our work, we had little confidence that we would provoke any more serious change. At the same time, we believed that the emerging electronic environment would allow, indeed challenge, us to take apart and reconstitute our assumptions as to who could ask what questions or pursue which topics, and in so doing contribute to the reinvention of our field. We believed that the long term consequences of digital tools would be revolutionary but we felt that we needed first to address concerns of the traditional print world: the two existing needs to which we felt we could contribute were the need for more linguistically sophisticated tools for those working with Greek texts and the delivery of data — especially pictures — which were expensive to publish in conventional publications and often impossible to track down in physical libraries.

Building any prototype digital library, however compact, is a challenge. Students of ancient Greece — or of any other culture, ancient or modern — need a vast information space if they are to find evidence to support their ideas or to gauge the innumerable forces at work in the literature, history, art and society of the time. We knew from the outset that a certain threshold of coverage was necessary before the database would support inquiry into questions of sufficient breadth and complexity for us to see what impact the new technology would have. Perseus 1.0, published in 1992, was thus a prototype of the real system: it taught us much about what was needed and it supported a number of scholarly projects[8] (indeed, the philological tools in Perseus 1.0 already provided a key research tool for my own research on Greek historiography). Perseus 2.0, which Yale will publish in 1996 (nine years after development began) and which has already begun to make its presence felt on the World Wide Web, contains the core materials that we envisioned in 1987. Our work to date thus completes a preliminary stage in the larger research agenda: Perseus 2.0 (and subsequent versions of Perseus) have finally reached a scale where they can provide a satisfactory laboratory with which to study the impact of this medium on our work as humanists.

The long term results are therefore only now beginning to take shape as technological change exerts increasing influence. Nevertheless, the broad outlines of change point, in our view, in a general direction that we had anticipated from our earliest work. The practice of classics (and surely of all scholarly disciplines) has begun to undergo profound change: the book (which, as the codex, was an invention of later Roman antiquity[9]) has already begun to share center stage with electronic documents; electronic documents tightly coupled with links to one another blur the distinctness (itself artificial) of the printed artifact, thus challenging the shape of traditional publications; new forms of encoding and disseminating knowledge have begun to appear (e.g., in Perseus a rule-based system for Greek morphological as opposed to traditional descriptive grammar,[10] the first glimmerings on the horizon of virtual reality as a means to represent three dimensional art objects or reconstruct sites); the internet has now established an entirely new pace for development and publication.

Nevertheless, if the continuity which linked professorial practice from c. 1880 (the rise of modern research universities in the United States) to 1980 (when computers first began to emerge as standard tools of classical scholarship) has begun to vanish, everything might change, but everything would also, in an important sense, stay the same. Our fundamental goals — reconstructing, indeed breathing life into, a vanished world, establishing a curriculum of study that would engage the imaginations and intellects of learners from childhood through adulthood, holding up the past as a mirror to the present — only grow stronger. Despite the

---

[7] http://csdl.tamu.edu/DL95/papers/nuernberg/nuernberg.html.

[8] Perseus has been under constant scrutiny and evaluation since it first began to make prototype materials available: Neuman, 1991, Samberg, 1993, Marchionini and Crane, 1994. Initial evaluation was funded by the Annenberg/CPB Project and by the office of Derek Bok, then president of Harvard University. The Fund for the Improvement of Postsecondary Education has provided funding that allowed colleagues at seven different institutions to participate in an extended evaluation from fall 1993 through spring 1996. Gary Marchionini of the College of Library Science at the University of Maryland has supervised the external evaluation team since the beginning; the most recent evaluation report is available on our Web Server: http://www.perseus.tufts.edu.

[9] O'Donnell, 1992.

[10] Crane, 1991.

challenges and pressures of modernity, emerging technologies provide us with an opportunity to realize more fully the larger goals that scholars have pursued since systematic study of the Homeric poems began in the fifth century BCE.

Our efforts and what we have learned constitute the main focus of this paper, but our work, which had its earliest beginnings in 1982, represents a second generation of effort within classics. Indeed, within the field of classics Perseus was, already in 1985, planned as a second generation digital library system. Before turning to our own work, we need first to clarify the nature of classics as a field and of those who pioneered the application of computing to our work.

The Perseus Project Digital Library on Ancient Greece

Our work on the Perseus Project grows directly out of experiences with the *Thesaurus Linguae Graecae* (TLG), a database now containing virtually all Greek texts surviving up through 600 AD, when it released its first materials on magnetic tape in 1982.[11] The author, then a graduate student at Harvard and now the Editor in Chief of the Perseus Project, had been primarily responsible for developing a Unix based full-text retrieval system (running at the time on a DEC PDP 11/44 with an ancient version of Unix) that could perform the standard information retrieval functions (locate words or phrases rapidly) or allow users to call up source texts by "chapter and verse," the conventional citation schemes used by classicists (and, in many ways, the intellectual precursors of Universal Resource Locators): e.g., "Thuc. 3.43.2" describes "Thucydides, *History of the Peloponnesian War*, Book 3, Chapter 43, Section 2." Other long term contributors to the project had also begun work on parallel lines at this period. In effect, we had begun to work with a digital library that contained many of the core materials for our field.

Experience with the TLG raised questions as to how broad a selection of material we could collect. The Greek source texts in the TLG constituted one crucial dimension of our field. We began to wonder what would happen if we could collect not only Greek texts, but all categories of information, from English translations to architectural plans.

Early planning for what would become Perseus began in the summer of 1985: an equipment grant from Xerox Corporation gave us access to Notecards, one of the first major hypertext systems. From September 1985 through May 1987, a project team and our initial assumptions took shape. Both the project team and assumptions of 1987 have proven remarkably resilient. Eight years of continuous development have led to many refinements, but the fundamental vision with which we began has remained.

Applying the taxonomy suggested by Nürnberg (1995), we decided to concentrate on the following:

Data: We needed a great deal of raw information. The TLG had done the ground-breaking work, placing on-line most surviving Greek texts, and we were thus able to concentrate on supplementary materials. Where the TLG provided Greek texts from Homer through 600 CE, we concentrated on providing English translations of these texts, a Greek-English Lexicon, glossary entries, essays on various authors, various notes to individual texts and other basic supporting materials. Ultimately, we assembled 3.4 million words of Greek texts along with accompanying translations (c. 67% of all Greek source texts and their translations surviving up through 300 BCE) as well as the standard 35,000 word teaching lexicon (the Liddell Scott *Intermediate Greek-English Lexicon*).

At the same time, we intended from the start to collect as much visual material as possible. The economics of print made the publication of pictures expensive and had thus, in our view, seriously damaged the study of Greek — indeed, of any — culture. Coffee table books with a few dozen high quality color prints provided a problematic infrastructure that prevented scholars from fully integrating the art and archaeology of the ancient world into their work. To combat this, we included in Perseus 2.0 more than 500 plans and drawings and 24,000 color images. Even this represents no more than a small start, but it still constitutes the most comprehensive single publication on this subject with which we are familiar.

Similarly, maps have played an important role throughout the project. Early in the project, Neel Smith began translating print maps into electronic form and developing ways to plot sites from database coordinates such as Latitude/Longitude or UTMs: the goal was to be able to apply a database of coordinates to any map. Ultimately, this effort developed into a project of its own. Working with GRASS, a Geographic Information System, and various source materials (including the Digital Chart of the World and Landsat data), Smith contributed c. 1,000 maps of the Greek world to Perseus 2.0.[12]

Metadata: Because classical texts are canonical and have been studied for many years, classicists have access to considerable "metadata." We entered major indices to Herodotus' *Histories*, Pausanias' Description of Classical Greece, and the ancient handbook of Greek mythology attributed to Apollodorus. These were merged and served as the basis of what we (somewhat hopefully) called our encyclopedia: c. 7,000 items, mainly short glossary identifications of people or things interspersed with a few essays on major topics, with most entries containing at least one link into materials within the database. In addition, we developed keywords that we then applied to the c. 2,000 art objects in our database.

The more we worked, the fuzzier the category of metadata seemed. Graphical interfaces to site plans would, for example, also fit under this category. We can plot c.

[11] Berkowitz, Squitier and Johnson, 1990 contains a brief account of the TLG; information on this project is available at http://www.uci.edu:80/~tlg/.

[12] For an overview of Smith's on-going work with GIS systems, see http://perseus.holycross.edu.

2,000 sites on our database of maps, but even these points on a map serve as links to other data, since users can move (more or less directly) from a point on a map to more information about that site. We also created hundreds of links between digital views of a site and positions on various plans. The site plans in turn have links to information about individual buildings, these building entries are likewise potential links into a database of ancient architecture. Even the Historical Overview of Ancient Greece, written by Thomas Martin for the Perseus Project, itself constitutes a major element of our metadata, since it is designed to introduce users to the database as well as the culture and to this end contains hundreds of links.

Library Processes: Most of our work on Perseus emerged from our experiences in developing full text retrieval tools for the TLG database of Greek texts. Information retrieval and the transformation of physical library processes have, in many ways, constituted one of our primary interests in the project. We have spent much of our time applying digital library processes to static data (such as the site plans mentioned above) so that they could serve also as metadata, leading users into other elements of the database (e.g., from plans to digital images illustrating those plans).

New Digital Objects: More can be Different

New digital objects are taking shape that are at least as different from their print counterparts as movies are from plays, and identifying these as they emerge is a major challenge for any digital library designer: a library system that cannot smoothly accommodate these objects will not last long. In some cases, these new digital objects are strikingly new: hyper novels, with their lack of linearity and protean form, cannot be confused with traditional books,[13] while the Virtual Reality Markup Language (VRML) promises to allow users on the World Wide Web to move through a virtual space.

Even without considering such purely digital technologies as three dimensional walk-throughs (a natural area of development for us) or tightly coupled links between text, sound and video (not as immediately applicable for students of ancient culture), the translation from print to digital medium transforms objects: simply being able to search a 500 page manuscript can represent a huge step forward. The ability to create links between a static plan and still images combined the two into a newly interactive system for browsing. By entering a Greek-English Lexicon we found that, by allowing others to search the English definitions, we had almost by accident created a rough English-Greek lexicon that, for all its coarseness, was in some ways more useful than any print counterpart.

Consider, for example, a case where change in a single dimension (quantity) leads to qualitative change. In the case of still images, we found early on that the electronic medium allowed us to view each electronic image as a modified translation of film: even now, as we begin work with digital cameras and leave film behind, the process of shooting individual photographs (at least for our purposes) has not changed. The ability to collect, manage and deliver thousands of images cheaply has, however, completely transformed our idea of what an electronic archive should be. Even when the "books" (in this case, images) remain much the same, the "digital library" is completely different.

A great deal of work has gone into developing image standards for digitizing microfilm[14] or making accessible fragile and hard to read documents, such as Greek papyri,[15] available to a wider audience. It rapidly became clear to us that existing source materials — large scale film transparencies, 35 mm slides, black and white prints, maps, drawings etc. — were not sufficient. Museum photographers, working with print publication in mind would, for example, regularly take small number of black and white pictures to illustrate an object. Photographic archives have thus adapted themselves to, and enshrined, the limitations of print publications, where any images are expensive and where, in many cases, economics rule out color pictures altogether.

Even in the earliest stages of our planning, when videodisk was the only viable technology for distributing still images and digital imaging was not an option for us, it was clear that the electronic medium differed fundamentally from print. A videodisk could store 54,000 color images on a single side — two orders of magnitude more than any normal publication, however lavishly illustrated. The individual frames on the videodisk had limited resolution (although, to be fair, these video images, though grossly inferior to slides or high quality prints, were comparable to most images printed in actual books). Nevertheless, the quantitative increase in storage capacity was enough to inaugurate a qualitative revolution in what purposes a "publication" could serve. In some cases, we found that complex objects, such as large Greek vases, could easily justify over one hundred separate pictures ranging from overviews to details of individual figures. For three dimensional objects, the videodisk was especially useful, since many medium resolution views were often more useful than a smaller number of the high quality color transparencies.

We found the same principles held true for site photography. Image archives would frequently contain pictures of a building from several key views, but rarely would the photographer turn his or her back on the building and shoot the view from the building, even though the physical setting — its location on a hill, for example — is crucial to sites such as the Sanctuary of Hera on the edge of the plain of Argos or the spectacular Sanctuary of Apollo at Delphi. Early on we collected motion video of several key sites, and we found that pans

[13] Moulthrop, 1991.

[14] See, for example, the CLASS project at Cornell: Crocca and Anderson, 1995: http://csdl.tamu.edu/DL95/papers/crocca/crocca.html; for a 1993 summary, see: http://www.xerox.com/PARC/dlbx/class.html.

[15] See, for example, the Duke archive: http://odyssey.lib.duke.edu:80/papyrus/).

and zooms conveyed a powerful sense of space that still images could not. Nevertheless, video was expensive to edit and, more importantly, consumes large amounts of storage. Even a series of ten-second pans and zooms eats up a videodisk or CD ROM very quickly, and we needed to cover a great deal of material in our database. To augment the video, we began experimenting with multiple exposure photography: the photographer would shoot six to eight pictures from a single strategic point, turning a bit with each and thus providing a 360 degree view. Similarly, we found it helpful to string these views together in linear walk-throughs, thus imitating the now venerable technique which the Media Lab employed in the Aspen Project.

Where we had planned to work with existing photography, we soon found that it was less expensive and more effective to hire our own photographer, who could collect images of consistently high quality and who would, even more importantly, shoot photographs optimized for the electronic world. Maria Daniels, who joined the project in 1989, was able to master the strengths and limitations before us, in the end taking roughly half of all the 24,000 images in Perseus 2.0.

Two major conclusions from our image collection are particularly relevant here:

1) None of the image archives that we consulted can support high-level digital image archives. Large-scale color transparencies may work well with two dimensional painting, but no archive had anything resembling the detailed color coverage that we needed to represent three dimensional objects such as sculpture, painted vases, or coins. Digital image archives that simply translate existing pictures into electronic form are the modern equivalent of early films, which planted a static camera in front of a scene and turned the movie into a grainy, silent, black and white stage play.

2) We understood from the start that our efforts were aimed at a medium in rapid transition. However hard we worked, we knew that we were groping in the dark and that our efforts might not meet the needs of imaging standards that would emerge in the future. A three dimensional painted Greek vase should obviously be represented as a database of 24-bit deep points located in an X, Y, Z coordinate scheme rather than as a patchwork of two dimensional images. Similarly, our still images of site photography allowed us to construct, in a very simple format, virtual tours — site plans with links to still images illustrating the view from various points — but this was clearly just a temporary substitute for a real-time walkthrough of the virtual space. In fact, it does appear that our investment will prove worthwhile: Apple's Quicktime VR, for example, has opened up the possibility of translating our two dimensional imagery into three dimensional space. While it is unlikely that the results will match those that we will acquire with objects and sites after we have had extensive experience with constructing virtual objects and spaces, the images that Maria Daniels collected between 1989 and 1992 are far better suited to this new environment than any other source materials that we encountered in North America or Europe.

Platform Independent Formats for Archival Data

A Clear Win: Film over Video. The most successful decision — and one that drew substantial criticism from some of our advisers at the time — was our policy of collecting, wherever possible, 35 mm film images. Not only did videodisks remain the only viable distribution medium for our digital materials for the first years of our work, but videodisks derived from slides had, in our experience, inferior image quality to videodisks mastered from 1" or even 3/4" videotape. For the first several years of development, the 35 mm slides not only cost far more than videotape, but they then had to captured by a frame grabber and in the end produced an inferior videodisk as an end product. Nevertheless, the initial investment has proven wise, since already in Perseus 1.0, the digital images — although smaller than, and indeed derived from, their videodisk counterparts — have proven more useful in many environments: because they reside on the same CD ROM as the rest of the data, they require no special hardware and can even be delivered via a network. With Perseus 2.0, where the digital images are derived from the slides and not from the videodisk, the digital images have now overtaken their video counterparts in every category.

At present, we are preparing to move entirely from film to digital, collecting all new images in a digital format and bypassing the film process altogether.

Right Reasons, Questionable Choice: Postscript for both Drawings and Architectural Plans. When we first began work, most digital images were still 1-bit deep bitmaps and our original delivery software, Apple's Hypercard, is still, after eight years, optimized for bitmap images. Nevertheless, it was clear from the start that we did not want to collect a database of bitmaps. We chose to encode our images in a higher level format that was output and system independent.

Given the hardware and software at our disposal in 1987, Adobe's Postscript seemed the best solution. We invested considerable labor in drafting scenes on famous Greek vases: being line drawings full of curves and with only a few colors, these were ideal for the first version of Adobe Illustrator. While the results were spectacular, the process was labor intensive and soon gave way to simple photography. Drawings of site plans were also beautiful, but Postscript was not designed for such tasks. We have been able to export individual Postscript plans into other formats better suited to architectural design, but the process is not automatic, and we would use some other format for these purposes if we were starting over.

Big Costs, Huge Potential, Growing Benefits: SGML Encoded Texts. When we began work in 1987, the obvious format for Greek texts was Beta Code: a scheme that the TLG developed to encode not only the Greek language but the basic page layout of printed Greek texts. Conversion from Beta Code to HTML is relatively straightforward (an observation confirmed by our experiences in developing a WWW interface for texts encoded by other projects in this format).

Elli Mylonas, who supervised the development of all texts in the Perseus database, was, however, already familiar with SGML and its potential benefits. We therefore chose from the start to encode all of our texts in SGML (and with an eye to the emerging Text Encoding Initiative SGML guidelines — although these only emerged as our own work developed).

The choice of a TEI-compliant SGML format represented a major commitment to long term usability. If we had concentrated on surface structure (i.e., this string is bold, this paragraph is centered) and ignored semantic structure (i.e., this string is the name of a speaker in the play, this paragraph is the title of the play), we could have given classicists more texts years earlier and with *no immediate loss in perceived functionality*. In neither Perseus 2.0 nor the evolving Web version of Perseus can users at present do anything with our SGML texts that they could not have done if we had converted them into Beta Code (or, using present schemes, HTML).

There are two reasons for this. First, the delivery software that we chose cannot handle SGML: in creating the Hypercard version of Perseus 2.0 we spent much of our time figuring out how to throw information out of our texts so that they would be usable in a very simple environment. Now, as we move to new platforms, more sophisticated software, such as Dynatext and Pat, does exist, but, although these tools would allow us to perform many dazzling operations, it is not at all clear that any software would make many of our texts in their *current* SGML form radically more useful.

This is not, however, an argument against SGML or the work that has been done so far. Our SGML encoded texts have simply not completed the transition from page images to fuller, semantically deep markup. We have made some progress. On the one hand, we have, for example, tagged poetic texts according to meter, and we could thus compare the language of the spoken and sung portions of a Greek play. We have marked up explicit quotes and we could thus automatically study the language of Plato by filtering quotations of Homer that introduce archaic or poetic elements into the text.

At the same time, however, our texts generally refine rather than augment the information which typography conveys in the print originals. Consider the text of Thucydides. His *History of the Peloponnesian War* consists in large measure of narrative exposition (who did what, when, and why) and of reconstructed speeches, presented as direct quotes and attributed to various historical figures. The narratives and speeches have wholly different styles and indeed appeal to very different audiences (the narratives being crucial for ancient historians with political philosophers paying far more attention to the speeches). Those who have studied Thucydides closely know that the language of the two styles is quite different and that many terms common in one style will show up rarely or not at all in the other. Any text search should not only return the aggregate number of times that a word or phrase appears in Thucydides as a whole but also its appearances in the narrative vs. speeches so that the researcher can rapidly see whether the subject of the search is biased one way or the other. Indeed, the dichotomy of narrative vs. quoted speech is sufficiently important that the print lexicon of Herodotus (Powell, 1966: originally compiled by hand in the 1930s) specifically marks citations to quoted speech.

Our SGML encoded version of Thucydides does not, however, distinguish between quoted speech and general narrative. In part, methodological problems make this distinction a bit messy because Thucydides occasionally paraphrases without quoting. Nevertheless, the main problem is that the typographic source for our Thucydides text does not highlight the difference between narrative and speeches. By contrast, the page layout — as generations of exam taking students have learned — allows readers to distinguish between spoken and sung portions of a Greek play, while block quotes in Plato also stand out because they appear, in modern texts, as indented paragraphs. The page layout almost demanded that we explicate the underlying semantic distinctions such as spoken/sung or text/block quotes.

It is easy enough in the case of Thucydides to handle the major speeches: we can look up a list of speeches in Thucydides and then tag them in our electronic text. Failing that, we can scan for beginning and close quotes to find speeches. In either case, the added tagging requires modest labor, but it does require labor: someone who understands the text must add to its "information vector" a dimension that is not apparent from the typographic style. There is nothing particularly controversial about the speeches: an editor is not overstepping his or her role in marking them (or even in deciding where to draw the often fuzzy line between narrative and paraphrased text). The problem is that this task requires an extra level of expertise and it thus slows down the text entry process much like a disk access or the activation of a distant URL slows down a computational task. And in some cases, the labor can be quite large. The language used by different heroes in the Homeric *Iliad* differs subtly and repays close study,[16] but our source text for the *Iliad* does not put quotation marks around quoted speech: virtually no typographic clues mark the hundreds of quoted speeches that pervade the text. To separate them out, someone has to go through more than 15,000 lines of poetry.

In short, end users of Perseus cannot yet see much, if any, of the advantages of the TEI-compliant SGML format — indeed, they have, if anything, paid an opportunity cost, because we could have delivered more texts sooner had we adopted a simpler format. Nevertheless, our investment in the TEI-compliant SGML format was clearly correct. It has vastly simplified the task of scholars who will now create a new generation of texts with more sophisticated markup. Our experiences with these SGML texts will allow them to determine more clearly what will and will not be useful, while it will be much easier to augment texts already in SGML form than to go from a weaker, "page image" format.

If, however, we consider the project as a whole and our current development processes, the SGML encoding has already begun to pay dividends because it allowed us to

---

[16] Martin, 1989.

produce HTML versions of our texts much more easily than would otherwise have been possible. As we add more encoding and provide ever more finely honed access to our textual materials, the value of this complex, but powerful, format will only increase.

Digital Libraries as Catalysts for Change

From our perspective, digital libraries are most interesting when they allow us to rethink the functions and audiences of traditional publications. Already authors in ACM journals, for example, adapt their publications to serve the creation of metadata by suggesting their own keywords. But how far will this process ultimately go? To what extent will specialists in a given domain rethink and redesign their publications to make them fit more dynamically into an electronic world? Our experience represents one end of the spectrum: ever since David Packard generated a concordance to Livy in the 1967 and then, a few years later, developed an entire operating system (with its own modified microcode optimized for rapid text searching), classicists have generally designed their own tools. Most progress in our field has come when classicists familiarized themselves with the technology. There are obvious drawbacks to such a tradition, but one benefit has been that the technology has led some members of the field to begin rethinking what constitutes the basic document types which underlie and indeed constrain our work.

Consider, for example, a publication about the Greek historian Thucydides. A specialist journal may expect the author to grapple closely with the views of many different scholars, taking into consideration the problems raised by textual variants or other abstruse factors. A university press may be more concerned with selling as many copies as possible: the detailed argumentation that gives a publication depth may drive away general audiences and will surely cut into the number of professors who assign a book to their students (the promised land of the university press publication!). Many arguments derive from a careful consideration of language and its meanings, but the last thing that a publisher wishes to see are large quotes of Greek.

There is, however, no reason that a well-designed electronic publication cannot personalize its appearance for different audiences. Tens of thousands of high school students read the *Oedipus Rex* of Sophocles each year in English translation. What better contribution could a scholar make than an article which could, in its most general form, provide a clear, but vivid argument to the tenth grader but which, if unraveled, could provide the rigor demanded by the most crusty specialist?

Such personalization is, of course, not intrinsically new: publishers have long produced separate editions for different audiences. Nevertheless, the abridged edition is usually a separate project and often quite hard to reconnect with the complete version. And separate editions are hard to maintain, tending to splinter the market.

Consider, as one example, the standard Greek-English Lexicon. Professional classicists still use the ninth edition of the Liddell Scott Jones *Greek English Lexicon* (Oxford 1940): almost 40 megabytes of data, this lexicon contains more than 100,000 entries and 500,000 citations to specific texts. Most students, on the other hand, work with the Liddell Scott abridged *Intermediate Greek-English Lexicon*, which contains c. 35,000 entries, contains references to authors (e.g. a meaning shows up in Thucydides) rather than full citations of specific passages (e.g. "Thuc. 2.37.2"), and is only roughly one fifth as big (7.5 megabytes). The large Greek lexicon had grown so unwieldy that the main body of the text has been untouched for over half a century: Oxford University Press is preparing its second supplement, a separate document containing corrections and revised articles rather than a revision of the whole. Moreover, the plates of the Intermediate Lexicon have not been changed since its first printing in 1888. The Intermediate Lexicon thus constitutes a largely mechanical abridgment of the scholarly Lexicon as it existed more than a century ago and fifty years before the ninth edition of 1940. This division between scholarly and pedagogical tools has had negative consequences for each. The research lexicon was too dense and complex for use by any but professionals, but its market is inherently small and resources have not been available to revise it for over fifty years. The student lexicon had a larger market, but it has languished untended for more than a century.

The simple translation from print to electronic form has undercut the distinction between student and research tool. We have now placed both the research and the student lexica on-line and have made them available via our web site to students. While our analysis is still in its early stage, the simplest changes — the ability to separate meanings with blank lines, the use of bold and italics, the ability to search a long dictionary entry for citations from a particular author — have transformed their habits. Usage strongly suggests that most of those consulting our Greek texts are students rather than faculty. When users look up the meaning of a word, they go, by default, to the research lexicon, and the system is thus biased in this direction. Nevertheless, users can easily switch to the briefer entry in the student lexicon. In practice, however, users have only chosen this option c. 20% of the time, even though the dictionary entries which are consulted tend be common words and thus the hardest to read.

It would be possible to adduce other instances in which the medium has reflected back upon, and changed the essence of, the content that it was designed to deliver: after all, the major premise of Marshall McLuhan's work was that the medium was not neutral, that it constrained, even if it did not define, the message. This ongoing revolution has begun to break down the barriers between computer scientist, librarian, and domain specialist: some have even suggested that training in computer science may become "more design-oriented and less preoccupied with low-level algorithms and control structures,"[17] even as those of us in the humanities begin to train our students in the technical possibilities and limitations that shape our ability to formulate questions and represent our conclusions.

---

[17]Wegner, 1995.

If fundamental document types break down, and the barriers between specializations weaken, however slightly and briefly, emerging digital libraries pose substantial and exciting questions: how do we define the underlying forces at work or generalize from the dizzying particulars of this messy new field if the medium itself undermines the structures upon which we base our analysis? Digital libraries inevitably reshuffle the deck, changing the set of tasks that any given individual can perform. How do we address traditional needs without reenshrining in fresh concrete the barriers and limitations familiar from print? These are questions which emerging "digital libraries" (of which Perseus 2.0 is one example) will, we hope, render more concrete as objects of inquiry.

References

Berkowitz, Luci, Karl A. Squitier, and William A. Johnson. Thesaurus Linguae Graecae: Canon of Greek Authors and Works. 3 ed. New York: Oxford University Press, 1990.

Crane, Gregory. "Composing Culture: the Authority of an Electronic Text." Current Anthropology 32 (1991): 293-311.

Crane, Gregory. "Generating and Parsing Classical Greek." Literary and Linguistic Computing 6 (1991): 243-245.

Crane, Gregory. "What is Perseus? What is it not? Comments on the Bryn Mawr Review of Perseus 1.0." BMCR 3.6 (1992): 497-502.

Crocca, William T., and William R. Anderson. "Delivering Technology for Digital Libraries: Experiences as Vendors." Digital Libraries '95: The Second Annual Conference on the Theory and Practice of Digital Libraries, 1995.

Fox, Edward A., et al. "Digital Libraries: Introduction." CACM 58.4 (1995): 22-28.

Marchionini, Gary, and Gregory Crane. "Evaluating Complex Systems and Processes: Methods and Results from the Perseus Project." TOIS 12 (1994): 5-34.

Martin, Richard P. The Language of Heroes: Speech and Performance in the Iliad. Ed Gregory Nagy. Ithaca: Cornell University Press, 1989.

McLuhan, Marshall. Understanding Media: the Extensions of Man. New York: McGraw Hill, 1964.

Moulthrop, S. Beyond the electronic book: A critique of hypertext rhetoric. Proceedings of the Third ACM Conference on Hypertext (Hypertext '91): ACM, 1991. 291-298.

Neuman, Delia. "Evaluation Evolution: Naturalistic Inquiry and the Perseus Project." Computers and the Humanities 25.4 (1991): 239-246.

O'Donnell, J. "St. Augustine to NREN: The Tree of Knowledge and how it grows." : NASIG, 1992.

Powell, J. Enoch. A Lexicon to Herodotus. 2nd ed. Hildesheim: Georg Olms, 1966.

Samberg, Mark. "Perseus 1.0." Computers and the Humanities 27.5-6 (1993): 409-415.

Wegner, Peter. "Multimedia Document Engineering for Nonmajors." DAGS 95: Electronic Publishing and the Multimedia Superhighway. Ed. James Ford, Phillia Makedon and Samuel A. Rebelsky. Boston: Birkhäuser, 1995. 25-27.

Wiltshire, Sian, et al. "Review of Perseus 1.0." BMCR 3.5.4 (1992).