Cultural Heritage Digital Libraries: Needs and Components

Gregory Crane

Perseus Project
Tufts University
Medford MA 02155 USA
gcrane@tufts.edu

Abstract. This paper describes preliminary conclusions from a long-term study of cultural heritage digital collections. First, those features most important to cultural heritage digital libraries are described. Second, we list those components that have proven most useful in boot-strapping new collections.

Introduction

This paper reports preliminary conclusions from the first three years of a five year project to develop a digital library for the humanities NSF IIS 9817484; [1-5].[1] We have established a set of testbed collections to complement substantial Greco-Roman materials that have been under development since 1987. We now have in place testbeds on early modern English, the history of mechanics, the history and topography of London, slavery and the US Civil War, and various collections on US History from the Library of Congress. These collections include multiple languages, spaces of various scales, and diverse classes of objects. Our goal was to create an unwieldy set of heterogeneous, in some ways incommensurable, collections that appealed to wide-spread, complex audiences with disparate, often competing interests. In so doing, we intended to pose issues of scalability and personalization that would have remained hidden had we explored the needs of a single area or audience. In the following two years, our research will focus on the interaction between front-end transactions and back-end structures, studying the differing strengths of the collections at our disposal and

[1] Primary funding for this work comes from the US Digital Library Initiative, Phase II: http://www.dli2nsf.gov. A substantial portion of the support from our work came from the National Endowment for the Humanities (http://www.neh.gov). International Digital Library grants from the NSF and *Deutsches Forschungsgemeinschaft* (DFG: http://dfg.de) and from the NSF and the EU have also contributed substantially to our recent work.

the varying needs of our audiences. This paper presents some preliminary conclusions and hypotheses about the needs and possibilities facing cultural heritage digital libraries [6, 7]. While our work considers representations of space, in two and three dimensions, and at various scales[8], from individual sites to global, and of objects such as art works or scientific instruments [9], this paper will concentrate on textual materials and the possibilities and challenges posed by human language technologies.

We are particularly interested in identifying the needs, present and potential, similar and dissimilar, of various communities within the humanities. We began our work three years ago by studying issues that different areas within the humanities did or did not share: students of the ancient world, for example, work intensively on relatively small, very fragmentary data sets in complex languages (e.g., Latin, classical Greek, Sumerian, Sanskrit); students of modern industrial cultures, by contrast, have extraordinarily detailed records and sources. Those working with pre-modern, typically sparse materials spend much of their time extrapolating from imperfect sources, while those working with recent and often vast data sources have a greater need to filter and visualize their data. Extrapolation and reduction are, of course, complementary processes and play a role in all research. The richness of information available about London, however, challenged us to develop visualization tools for space and time that we would not have otherwise pursued but that have proven powerful also for the Greco-Roman collection.

Many issues confront digital libraries in a variety of areas and, indeed, our growing interactions with the NSF National Science Digital Library (http://www.nsdl.nsf.gov) suggest to us that the interests of humanists, social scientists and natural scientists are converging. Although the emphases may differ, all the recommendations from the June 2001 Delos Digital Library brainstorming session in San Cassiano [8], for example, are relevant to cultural heritage collections. Reading support systems that help students read Greek and Latin texts have, for example, provided the foundation for services that support interdisciplinary researchers and undergraduates as they shift from textbook culture to real scientific literature. Such services, however, raise issues about how we encode data and structure the architecture of

our DL environment. Administrative guidelines complicate the role of humanists in a National Science Digital Library funded by the National Science Foundation. Nevertheless, humanists can play a vital role by bringing distinct perspectives that may complicate short-term goals but lay the foundation for a system that is in the long run more robust, general, and sustainable.

The paper has two main parts. In the first, we outline a set of overall issues that are, in aggregate, particularly important to cultural heritage collections. The second section describes what elements we have found to be useful in boot-strapping cultural heritage collections,

Characteristics of the Humanities

This paper reports preliminary conclusions from the first three years of a five year project to develop a digital library for the humanities NSF IIS 9817484; [1-5].[2] We have established a set of testbed collections to complement substantial Greco-Roman materials that have been under development since 1987. We now have in place testbeds on early modern English, the history of mechanics, the history and topography of London, slavery and the US Civil War, and various collections on US History from the Library of Congress. These collections include multiple languages, spaces of various scales, and diverse classes of objects. Our goal was to create an unwieldy set of heterogeneous, in some ways incommensurable, collections that appealed to wide-spread, complex audiences with disparate, often competing interests. In so doing, we intended to pose issues of scalability and personalization that would have remained hidden had we explored the needs of a single area or audience. In the following two years, our research will focus on the interaction between front-end transactions and backend structures, studying the differing strengths of the collections at our disposal and the varying needs of our audiences. This paper presents some preliminary conclusions and hypotheses about the needs and

---

[2] Primary funding for this work comes from the US Digital Library Initiative, Phase II: http://www.dli2nsf.gov. A substantial portion of the support from our work came from the National Endowment for the Humanities (http://www.neh.gov). International Digital Library grants from the NSF and *Deutsches Forschungsgemeinschaft* (DFG: http://dfg.de) and from the NSF and the EU have also contributed substantially to our recent work.

possibilities facing cultural heritage digital libraries [6, 7]. While our work considers representations of space, in two and three dimensions, and at various scales[8], from individual sites to global, and of objects such as art works or scientific instruments [9], this paper will concentrate on textual materials and the possibilities and challenges posed by human language technologies.

We are particularly interested in identifying the needs, present and potential, similar and dissimilar, of various communities within the humanities. We began our work three years ago by studying issues that different areas within the humanities did or did not share: students of the ancient world, for example, work intensively on relatively small, very fragmentary data sets in complex languages (e.g., Latin, classical Greek, Sumerian, Sanskrit); students of modern industrial cultures, by contrast, have extraordinarily detailed records and sources. Those working with pre-modern, typically sparse materials spend much of their time extrapolating from imperfect sources, while those working with recent and often vast data sources have a greater need to filter and visualize their data. Extrapolation and reduction are, of course, complementary processes and play a role in all research. The richness of information available about London, however, challenged us to develop visualization tools for space and time that we would not have otherwise pursued but that have proven powerful as for the Greco-Roman collection. Perceived initial differences of the London collection set us moving in a new direction but the results were in the end unexpectedly useful for the Greco-Roman collection.

Many issues confront digital libraries in a variety of areas and, indeed, our growing interactions with the NSF National Science Digital Library (http://www.nsdl.nsf.gov) suggests to us that the interests of humanists, social scientists and natural scientists are converging. Although the emphases may differ, all the recommendations from the June 2001 Delos Digital Library brainstorming session in San Cassiano [10], for example, are relevant to cultural heritage collections. Reading support systems that help students read Greek and Latin texts have, for example, provided the foundation for services that support interdisciplinary researchers and undergraduates as they shift from textbook culture to real scientific literature [on which, see below]. Such services, however, raise issues about how we encode data and

structure the architecture of our DL environment. Administrative guidelines complicate the role of humanists in a National Science Digital Library funded by the National Science Foundation. Nevertheless, humanists can play a vital role by bringing distinct perspectives that may complicate short-term goals but lay the foundation for a system that is in the long run more robust, general and sustainable.

The paper has two main parts. In the first, we outline a set of overall issues that are, in aggregate, particularly important to cultural heritage collections. The second section describes what elements we have found to be useful in boot-strapping cultural heritage collections, We then advance the notion of a corpus editor — a scholar with expertise in a particular area managing a corpus whose size is too large for manual methods of editing and who must therefore rely upon automated (and thus inherently imperfect) methods [2, 11]. This paper documents some of the concrete tasks that were required to develop a new collection of nineteenth century materials on slavery and the US Civil war. This study builds on results with collections on the Greco-Roman world and on the history and topography of London.

1. Historical data become more valuable over time — persistence is crucial: Cultural heritage digital libraries must aggressively address the problem of digital preservation. The problem is particularly serious for complex knowledge sources such as lexica or encyclopedias. Humanists may be less able than their colleagues to retrofit gigabytes of complex materials, but humanist reference works are used for decades, if not longer.
2. Access to the cultural heritage of humanity is a right, not a privilege: The record of human achievement is a public good and should be accessible to every citizen. At present, private corporations have undertaken the crucial task of digitizing some critical corpora and have produced intellectual gated communities. These electronic resources, tightly controlled and often priced in such a way as to guarantee a limited audience, restrict fundamental source materials to the same academic elites that had access to scarce print resources. A socio-economic infrastructure has thus begun to arise that imposes on the digital world limitations of print. We need economic models that do not replicate practices that isolate cultural heritage from the community as a whole. Governmental approaches are, however, also problematic, since governments may feel an obligation, explicit or not, to control their national image and impose restrictions on information.
3. Cultural heritage digital libraries must serve the needs of diverse audiences: Access to information is necessary but not sufficient. Customization is a rapidly growing field of inquiry. The system should adapt to the needs of its users, providing them with the information that they need to interpret new documents or topics, reducing,

insofar as possible, the friction of their movement through a digital library (e.g.,[12]). There are limits to this — as Euclid reportedly rebuked the first Ptolemy with the statement that there is "no Royal Road to geometry,"[3] some concepts are simply difficult. Nevertheless, a humanities digital library has a social obligation to support the development of complex skills, by a wide audience, over a long period of time.

4. The *documents within* cultural heritage digital libraries must serve the needs of diverse audiences: Humanists cannot simply rely upon elaborate technologies to enhance their contributions to society as a whole. The Internet already reaches a huge audience and could within a very near future saturate the households of the advanced countries. The Perseus Digital Library Website has, for example, emerged as a major distribution channel within classics and now disseminates up to 9,000,000 pages of data per month to an audience far beyond traditional academia. Humanists — especially those who participate in scholarly debates that span decades or more — must think carefully about how they will respond to this vast new and expanding audience. We need to ponder both the way in which we write and the questions that we pursue. Maintaining the status quo and dismissing this new audience is itself a strong, if problematic, response.
5. The library is a laboratory where reading is a primary exercise: To some extent, this is a superset of the customization problem. A great deal of DL research addresses the cataloguing problem. A digital library is a structured space that manages a large number of objects. The user searches through the DL to find objects of interest, but, once these have been found, many systems simply hand control over to the object and the user calls up a PDF viewer etc. Humanists often study texts, images and spaces in extremely close detail. Thus, the numbered citation schemes of computer science publications — which direct readers to a document as a whole — reflect a much less general attitude to textual reference: humanists are trained to cite precise pages and, when dealing with canonical documents, often cite individual lines or words. In this environment, the granularity is much finer and users need support with words and phrases as well as with documents as a whole. The implications are, however, profound for the scale and design of humanities DLs: when each word becomes a complex multidimensional object, density of data increases by several orders of magnitude. Cultural heritage materials raise challenges that go beyond those described in the literature about citation harvesting and linking from recent scientific publications [13, 14].
6. Digital objects and their components must be freely reusable: Simple access to information is not sufficient. We need complex documents that include and provide distinct visualizations of components from many sources, e.g. details from high-resolution images, clips of time-based media, tabular or graphic visualizations of data sets, quotations from larger works, and links from each inclusion to the source.
7. Standards/best practices must be descriptive rather than prescriptive: New publication series can impose guidelines on the form and structure of documents.

---

[3] Reported by Proclus in his description of Euclid: see http://ww.perseus.tufts.edu/cgi-bin/ptext?doc=Perseus:text:1999.01.0086&query=head%3D%232&word=Euclid.

The variations of historical sources can provide crucial information. We thus need to preserve, rather than eliminate, vagaries of spelling in early modern texts since these variations can provide important data about the compositional history of a given text [6, 15] (e.g., compositors often provided the actual spelling and uses of "do" vs "doe," for example, can help determine who is responsible for what. section of Shakespeare's First Folio. The need for prescriptive rather than descriptive encoding demands a consequently far more complex encoding scheme and software infrastructure. This requirement generates a need in turn for specialized viewers, which can, for example, filter and display very precise differences between editions. While the underlying ideas are similar to the well-known problem of versioning source code, a cultural heritage versioning system requires substantially more precision of reference and semantics: editors within the *New Variorum Shakespeare* series, for example, formally distinguish between "substantive" and "semi-substantive" changes to the text [16]. A versioning system must be able to manage a wide variety of such classes.

Building Cultural Heritage Collections

When planning for Perseus first began in 1985, we wanted to create a critical mass of information about the classical Greek world. While the *Thesaurus Linguae Graecae* [17] had already created a digital library of classical Greek source texts, we wanted to create an environment that contained every category of information about the Greek world: not only Greek source texts, but aligned translations, grammars, lexica, encyclopedias, reference articles, as well as maps, plans, pictures of places and art objects, catalogues, narrative discussions of art and archaeology., etc. Very little existed and we needed to create a self-standing, heterogeneous environment with which to experiment. We had a full-time photographer who took original images in dozens of museums across North America and Europe because few photo archives had the detailed, consistent photographic coverage that videodiscs, with room for 54,000 still color images on a side, could deliver. We digitized texts, drafted plans, commissioned articles, and addressed in-house as many tasks as we could. The results were satisfactory. Our goals were to create a corpus that was (1) large enough to support useful tasks of various kinds and (2) not tied to any one system. In the mid 1990s, we were able, with minimal effort, to migrate the Perseus DL from a *Hypercard* delivery environment to the Web and are prepared to shift data collected since the 1980s to new systems in the future. The Greco-Roman materials in Perseus remain

popular, accounting for roughly 85% of the 26,000,000 pages that we served from February through April of 2002.

The rise of the Web and, more recently, of the *Open Archives Initiative* (http://www.openarchives.org/) has radically changed the information environment. If we were beginning Perseus now, we would clearly not pursue the same strategies that we adopted in the 1980s. Nevertheless, the independent work that we did many years ago continues to prove immensely useful. In part, this reflects the fact that we have control over a number of digital objects on which we can, within some limits, freely experiment. Third parties are often understandably unenthusiastic about others modifying their carefully designed data-structures.

Collection Overview: We had the resources when first developing the Greek collection in Perseus to commission from the ancient historian Thomas Martin a new book-length overview of Greek history and culture. Professor Martin produced a work that also appeared in print form [18]. The electronic version, however, was designed as a hypertext, with many cross-references complementing a clear hierarchical structure. More importantly, the electronic overview contained thousands of links to primary materials on which interpretive statements were based. Where the print version was designed for the isolated reading typical of a book aimed at a broad market, the electronic edition was designed to guide readers to the primary sources. Furthermore, while many of the links were deterministic (they pointed to particular passage or objects), some were dynamic queries: e.g., "search for women and slaves in the Greek collection." The results of these dynamic queries are different now from what they were when the *Overview* was first published on the Web and will continue to change, as the Greco-Roman collection evolves.

The *Greek Historical Overview* has been the most popular single work in the Perseus Digital Library. The *Overview* represented a broad synthesis of the field of Hellenic studies and drew upon the full experience of a senior faculty member. It thus constituted a major investment of time, money and support, but the results have more than justified the costs. We have created more modest introductions to other Perseus collections, but the Greek overview remains a key element and

a potentially important case study in electronic publication. Not only does it contribute to the collection, but increased exposure of the web publication and the links binding text to sources have increased the audience that it has reached and the intellectual contribution which this work has been able to make. Nor have sales of the print version suffered (despite early concerns by the publisher).

Images: Digital cameras now allow novices to produce images that are useful in many contexts and institutions have begun using new technology to document their collections. The OAI is well suited to disseminating images and we are already harvesting metadata records of images that complement our collections. We no longer, for example, maintain a full-time staff photographer in Perseus nor do we see the need to provide the sorts of encyclopedic coverage that we attempted for the original Greek Perseus.

Nevertheless, the photography that we commissioned in the first ten years of the project remains a core resource. The image quality and, more importantly, the coverage are consistent: these photographs were taken to support digital publication and we assumed that we would be able to publish as many images of an object as we saw fit. In photographing Greek vases, we shot overviews from multiple angles and close-ups of individual scenes, figures and significant details — for some particularly complex vases, we shot almost two hundred individual views. The aggregate photographic coverage provides an immense amount of information and raises interesting opportunities for data-fusion techniques to stitch the disparate views into a single massive database.

Narrative texts: In this context, we define narrative texts as documents with relatively simple typographic/organizational structures (e.g., chapters/sections) that lend themselves to OCR and rapid tagging. Many thousands of public domain literary texts are now freely available, often from multiple sources, on the Web. Third party texts are, however, problematic. The quality of data entry is uneven. The bibliographic source may not be listed and even such basic citation schemes as page breaks may be lacking. Even when texts that are well edited and encoded in SGML/XML, the publicly available Web version may be informationally diluted HTML and servers may (like the

Perseus Digital Library) only provide subsets of a text at a time. While third party sources are promising, they do not always remove the need to (re)digitize large local collections.

In developing our collection on London, we have chosen to enter a small number of canonical works available from other sites (e.g., Dickens' *Little Dorrit*) for experimental purposes but will rely on third parties for broader coverage. It remains to be seen how important canonical citation schemes are for most narrative texts associated with the London corpus.

2D and 3D models: While this paper focus on language issues, any mature digital library must develop a strategy to integrate 2D and 3D spatial data with textual and other materials: digital libraries provide a space in which user can theoretically move back and forth between virtual spaces, textual sources and quantitative data. Thus, users moving through a virtual London or Republican Rome should be able to ask questions such as "what is the architectural style of this building?" Such integration is hard to achieve, however, if models are developed in isolation and only with a view to generating imagery. We have long created vector based 2D models of archaeological sites and have in the past several years begun developing 3D models as well. While the technical tools for 3D modeling are well established, the scholarly conventions for vector models are still very much in flux. Where industrial developers may focus on photorealistic modeling, academics need an environment in which to critique and analyze models. A scholarly system should be able to provide the "state plan" (e.g., our evidence, whether archaeological excavations or contemporary observations of a historical space), multiple reconstructions along with the evidence on which those reconstructions are based, and details about individual elements of a space (e.g., point to a "Corinthian column" and locate similar Corinthian columns from other models). Such functionality requires data structures and labeling that are not typical of most professional drafting. The energy that we put into digital photography has now, in effect, shifted to modeling.

On the other hand, we are developing models of historical spaces as examples and case studies, designed to educate ourselves and to

provide insights to others on how such data might function as part of an integrated digital library.

Geospatial data: Geospatial data have been crucial to our work from the beginning, but we have always relied upon third party gazetteers (such as the *Getty Thesaurus of Geographic Names* and the newly released gazetteer from the Alexandria Project) for our base data to which we have added supplementary information. We have accomplished a substantial amount by combining large third party datasets with modest human data collection, but the improved accuracy of GPS data and the ubiquitous availability of handheld GPS units open up immense new possibilities for collaborative collection of point data. We do not have in place the collaborative infrastructure needed so that the archaeologists who fan out across ancient sites all over the world each year can relay GIS data to a central repository. Such an infrastructure is a major desideratum for cultural heritage collections. We are moving in that direction, as are other efforts such as the *Electronic Cultural Atlas Project* at Berkeley (www.ecai.org).

Lexica: In 1987, we digitized a Greek lexicon before we had digitized any Greek texts — a prioritization that proved extremely useful and which we continue to follow. At present, we are digitizing two dictionaries of classical Arabic to spear-head a movement towards an Arabic collection.

Starting with a dictionary can seem problematic. Dictionaries tend to be large (the major Greek [19] and Latin [20] lexica were 35 megabytes each), to have complex formats that do not quite capture the logical structure of the entries, and to defy efficient OCR (although promising work is going on at the University of Maryland on OCR and analysis of dictionaries:[21]). Dictionaries are thus far more expensive than source texts, since they may need to be hand keyed, with surcharges for unfamiliar writing systems such as Greek or Arabic. Converting a 35 megabyte print lexicon to a useful electronic resource that can drive a morphological analyzer may cost as much as a double keyed 200 megabyte library of source texts. Lexicographers aside, most see dictionaries as a means to read other materials. Beginning a digital collection in a new language by lavishing time and money on a lexicon is thus not an obvious decision.

Nevertheless, the dictionary is a crucial starting point precisely because it is expensive, difficult and powerful. Our on-line dictionaries have provided a foundation for many subsequent services and add value to every text linked to the system. We mine the dictionaries for their morphological information, use these data to drive a morphological analyzer, and then provide dictionary look-up and lexically based search services. In English this would be equivalent to (1) being able to click on "were" and calling up the entry for "to be," along with information about the form "were" and (2) searching for "to be" and retrieving "were." In highly inflected languages such as Greek, Latin, and Arabic, morphological analysis can be extremely complex but the resulting services correspondingly important. Anecdotal reports suggest that students read Greek and Latin twice as fast using the dictionary links in the Perseus environment as when working with print — whether or not those figures are accurate, the perceived increase in throughput has attracted substantial use. The benefits of more powerful searching are harder to quantify but substantial.

Grammars: We have entered grammars for Greek and Latin. In theory, these grammars may provide data-sources to support syntactic and semantic analysis of Greek and Latin. Where we have mined the morphological data from the lexica, the syntactic and semantic information remains embedded in the texts and we have not been able to generate linguistically based services from these resources. These grammars do, however, contain thousands of precise citations of source texts. We convert these citations to bi-directional links, so that those reading a given text can see when a grammar comments on a particular passage. Although the grammars are, in their present form, cumbersome to browse, the bi-directional links generate substantial usage and have made them popular resources within the collection.

Encyclopedias: Dictionaries concentrate on semantics — the general meanings of words. Many texts contain references to particular people and places. We did not have access to a classical dictionary when developing Greco-Roman Perseus but we used glossed indices for key reference works to provide basic information about 7,000 mythological and historical figures. We supplemented this with roughly one hundred commissioned new articles on key authors, sites and topics. We then

used simple pattern matching to attach automatic links to Perseus texts. When the reader sees a reference to Alexander, a link appears that leads to descriptions of all the Alexanders about whom we have information, including Alexander the Great. We can compare the words around the Alexander in a given text to the language in the entries on the various Alexanders to determine which Alexander is probably meant, but even without such filtering, the lookup service became immensely popular and remains the most widely used function in the Perseus Digital Library.

Other Reference Works: Not all reference works are, like encyclopedias, organized with self-contained articles under discrete keywords. The London collection contains many guides — some many volumes long — describing the city. The organization is hierarchical and topographic: e.g., one section will cover Westminster and then follow the path of a hypothetical visitor. Sometimes individual buildings will have their own self-contained entries. In other cases, bold type or some other high-lighting will indicate how the text's focus shifts from one building/place to another. Tagging strategies can thus be crucial: a few hours can suffice to determine which italics phrases are keywords and which are foreign language quotes, regular emphasis, etc. Once keywords are identified as such, they can be used to generate automatic links and to lead readers to the relevant sections. It is easier to build this into the workflow for digitization than to retrofit dozens, if not hundreds, of such documents later.

All on-line data derives its value from the technologies that mediate between the user and the bits, but many of the most promising technologies require human mediation if they are to prove useful. The final section of this paper builds on the concept of a corpus editor, which we have introduced in earlier publications [2, 11].

Conclusion

This document has described some of the requirements and possible components that we have found to be important to cultural heritage collections. We have stressed those features that seem to us most distinctive but we have suggested that the particular needs of cultural

heritage collections only anticipate issues that other disciplines may confront as their usage of digital libraries increases. We offer one service as an example of such a convergence.

Much of this paper describes strategies to support reading. Morphological analysers link inflected forms to dictionary entries. Information extraction systems map words to particular things. The importance of language and historical context have led us to focus on services of this type at an early stage of development. We are developing a knowledge tracking system for language readers: the system keeps track of what textbooks and readings an intermediate language student has read. When the student requests a new text in the target language, the system identifies which words the student should know and which unknown words are particularly important, given their frequency in this document and the interests of the student.

In the foreign language scenario, we seek to track semantics, syntax, grammar, and references to particular people and things. Consider, though, the situation of the interdisciplinary researcher moving into a new field or the student moving from text books to real scientific literature. The same infrastructure aimed at reading support can track the technical terms that the researcher/student should already know and should learn. The infrastructure is the same — indeed, in a hierarchy of difficulty technical terms are easier to identify than encyclopedic references (e.g., “Mr. Smith,” “Springfield”) and much easier to identify than the particular meanings of natural language (e.g., bank as “river bank” or “money bank”). By attacking the harder foreign language problem we lay the foundation for a service that supports reading in the scientific and medical areas as well.

References

1. Smith, D.A. and G.R. Crane, *Disambiguating Geographic Names in a Historical Digital Library*. 2001, Perseus Project/Tufts University: Medford, MA. http://www.perseus.tufts.edu/cgi-bin/ptext?doc=2000.06.0012.
2. Rydberg-Cox, J.A., A. Mahoney, and G.R. Crane. *Document Quality Indicators and Corpus Editions*. in *JDCL 2001: The First ACM+IEEE Joint Conference on Digital Libraries*. 2001. Roanoke, VA, USA: ACM Press.
3. Crane, G. *The Perseus Project and the Problems of Digital Humanities.* in *Standards und Methoden der Volltextdigitalisierung*. 2001. Trier, Germany: Mainz Academy.
4. Crane, G., et al., *Drudgery and Deep Thought: Designing Digital Libraries for the Humanities.* Communications of the ACM, 2001. **44**(5).

5. Crane, G., D.A. Smith, and C. Wulfman. *Building a Hypertextual Digital Library in the Humanities: A Case Study on London*. in *JDCL 2001: The First ACM+IEEE Joint Conference on Digital Libraries*. 2001. Roanoke, VA, USA: ACM Press.
6. Furuta, R., et al. *The Cervantes Project: Steps to a Customizable and Interlinked On-Line Electronic Variorum Edition Supporting Scholarship*. in *European Conference on Digital Libraries (ECDL 2001)*. 2001. Darmstadt, Germany: Springer.
7. Brocks, H., et al. *Customizable Retrieval Functions Based on User Tasks in the Cultural Heritage Domain*. in *European Conference on Digital Libraries (ECDL 2001)*. 2001. Darmstadt, Germany: Springer.
8. Chavez, R.F. and T.L. Milbank. *London calling: GIS, VR, and the Victorian period*. in *7th International Conference on Virtual Systems and Multimedia*. 2001. Berkeley, CA.
9. Daniels, M. *Is bigger better? web delivery of high-resolution images from the Museum of Fine Arts, Boston*. in *Museums and the Web*. 2000.
10. Ioannidis, Y., B. Croft, and E. Fox, *Digital Libraries in the Future: A Grand Challenge Vision for the 6th Framework Programme*. 2001, DELOS: Network of Excellence: Pisa, Italy.
11. Crane, G. and J.A. Rydberg-Cox. *New Technology and New Roles: The Need for "Corpus Editors"*. in *The Fifth ACM Conference on Digital Libraries*. 2000. San Antonio: ACM.
12. Hong, J.-S., B.-H. Chen, and J. Hsiang. *XSL-based Content Management for Multi-presentation Digital Museum Collections*. in *European Conference on Digital Libraries (ECDL 2001)*. 2001. Darmstadt, Germany: Springer.
13. Bergmark, D. and C. Lagoze. *An Architecture for Automatic Reference Linking*. in *European Conference on Digital Libraries (ECDL 2001)*. 2001. Darmstadt, Germany: Springer.
14. Mahoui, M. and S.J. Cunningham. *Search Behavior in a Research-Oriented Digital Library*. in *European Conference on Digital Libraries (ECDL 2001)*. 2001. Darmstadt, Germany: Springer.
15. Hinman, C., *The printing and proof-reading of the first folio of Shakespeare*. 1963, Oxford,: Clarendon Press. 2 v.
16. Hosley, R., R. Knowles, and R. McGugan, *Shakespeare Variorum Handbook: A Manual of Editorial Practice*. 1971, New York: Modern Language Association of America. 143.
17. Pantelia, M., *The Thesaurus Linguae Graecae*. 2002. http://www.tlg.uci.edu/~tlg/.
18. Martin, T.R., *Ancient Greece : from prehistoric to Hellenistic times*. 1996, New Haven: Yale University Press. xiii, 252.
19. Liddell, H.G., et al., *A Greek-English lexicon*. A new rev. and augm. throughout ed. 1940, Oxford,: The Clarendon Press. xlviii, 2111.
20. Andrews, E.A., et al., *A Latin dictionary founded on Andrews' edition of Freund's Latin dictionary. Rev., enl., and in geat part rewritten*. Rev., enl. ed. 1955, Oxford,: Clarendon Press. 1 l.,xiv,2019.
21. Oard, D., *Multimodal/Multingual Tools*. 2001, DARPA ITO Sponsored Research. http://www.darpa.mil/ipto/psum2001/J293-0.html.

22. Knight, E.H., *Knight's American mechanical dictionary*. 1877, New York, Hurd and Houghton, Cambridge [Mass.]: The Riverside press. 3 v.
23. Wheatley, H.B., *London Past and Present: Its History, Associations, and Traditions*. 1891, London: John Murray.
24. Dyer, F.H., *A compendium of the war of the rebellion*. 1908, Des Moines, Iowa,: The Dyer publishing company. 1796.
25. Hirschman, L., et al. *Integrated Feasibility Experiment for Bio-Security: IFE-Bio, A TIDES Demonstration*. in *HLT2001*. 2001. San Diego, CA.
26. Association for Machine Translation in the Americas. Conference (4th : 2000 : Cuernavaca Mexico) and J.S. White, *Envisioning machine translation in the information future : 4th Conference of the Association for Machine Translation in the Americas, AMTA 2000, Cuernavaca, Mexico, October 10-14, 2000 : proceedings*. 2000, Berlin ; New York: Springer. xv, 254.
27. Peters, C., ed. *CLEF 2001: Cross-Language System Evaluation Campaign*. 2001: Darmstadt, Germany.
28. Radev, D.R., S. Blair-Goldensohn, and Z. Zhang. *Interactive, Domain-Independent Identification and Summarization of Topically Related News Articles*. in *European Conference on Digital Libraries (ECDL 2001)*. 2001. Darmstadt, Germany: Springer.
29. Wayne, C.L. *Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation*. in *LREC 2000: 2nd International Conference on Language Resources and Evaluation*. 2000. Athens, Greece.
30. Voohees, E.M. *Overview of the TREC 2001 Question Answering Track*. in *TREC 2001*. 2001. Gaithersburg, MD 20899: NIST.
31. Zaslavsky, A., A. Bia, and K. Monostori. *Using Copy-Detection and Text Comparison Algorithms for Cross-Referencing Multiple Editions of Literary Works*. in *European Conference on Digital Libraries (ECDL 2001)*. 2001. Darmstadt, Germany: Springer.