Towards a Cultural Heritage Digital Library

Gregory Crane, Clifford E. Wulfman, Lisa M. Cerrato, Anne Mahoney, Thomas L. Milbank, David Mimno, Jeffrey A. Rydberg-Cox, David A. Smith, Christopher York

Perseus Project
Eaton Hall
Tufts University
Medford MA 02155

E-mail: {gcrane, lcerrato, amahoney, tmilbank, mimno, cwulfman, yorkc}@perseus.tufts.edu, rydberg-coxj@umkc.edu, dasmith@jhu.edu

Abstract: This paper surveys research areas relevant to cultural heritage digital libraries. The emerging National Science Digital Library promises to establish the foundation on which those of us beyond the scientific and engineering community will likely build. This paper thus articulates the particular issues that we have encountered in developing cultural heritage collections. We provide a broad overview of audiences, collections, and services.

Keywords: automatic linking, collection development, document design, reading, browsing.

1. INTRODUCTION

The efforts of the Perseus Project are based on a strong and somewhat polemical premise: namely, that digital libraries promise new methods by means of which new audiences can ask new questions about new ideas they would never otherwise have been able to explore. While we are based in a university and we are products of US higher education, we see the peer-to-peer interactions between professional colleagues [1-3] and indeed the formal instruction of 18-22 year old students [4, 5] as instruments for a broader purpose. In one recent survey of 1,500 people, "two fifths … reported that they pursue a hobby or collection related to the past, and they spoke of those pursuits with words like 'love' and 'passion'" [6]. In the broadcast world, twenty million Americans watched Ken Burns' Civil War series [7], while the History Channel and high end series such as WGBH's American Experience draw passionate audiences. Tens of millions of Americans visited history museums in the past year. The engagement with the past may be light-heartedly intense (e.g., a fascination with 19th century railroads) or may confront us with our darkest nightmares (as with the Shoah project).

We see in digital libraries an environment that can break down the barriers between academia and broader historical discourse about the past. The Americans surveyed above reported feeling "unconnected to the past in history classrooms because they don't recognize themselves in the version of the past represented there" [8, 9]. Academic historians, by contrast, express frustration with popular histories (see, for example, the debate around Burns' Civil War [7, 10, 11]). Digital libraries can reinforce existing structures, providing ever more specialized data to scholarly elites and ever more edutainment to society at large. But they can also, if so designed, expose the specialists to the challenges of a far wider and more truly diverse audience than any we encounter in the academy, while providing the authors, producers, and harried developers of websites for popular audiences with a much richer foundation on which to build.

We have approached this broad challenge from a much more modest background. Most of us in our group are trained as classicists, and the Perseus Project concentrated on Greco-Roman antiquity for its first ten years of work (1986-1996). Several factors, however, inspired us to expand beyond our own initial field. First, we had brought the Greco-Roman collections to a reasonable level of maturity. Second, we realized that, unlike biology or physics, classics was not large enough to sustain its own specialized digital library infrastructure. Classical languages raise serious challenges in digital library design; nevertheless, the largest humanities communities in the United States work primarily with English and a few major modern languages. It became clear that, to ensure that our particular needs were not left out of consideration, we needed to share infrastructure with humanists in other disciplines and define our common needs and objectives. The same logic has drawn us into the NSDL and led to an NSDL services project for reading support [13]. Third, a classical digital library is a strategic resource, since Western education and culture stressed classical models and Latin remained a major vehicle of scientific, literary and cultural publication through the eighteenth century. No one can understand the intricacies and subtle needs for all the numerous domains within the humanities, but an ability to handle Latin in particular, multilingual documents in general, and a range of visual materials gave us the tools to undertake a range of projects.

We could not, of course, create a digital library for all of cultural heritage – such a project would be vast and would have to be global in scope. Our Greco-Roman work provided a start for far more areas of Western culture than we could ever hope to explore. We have chosen a number of areas in which to develop collections. Our decisions have reflected difficult cost/benefit tradeoffs and have been controversial. The issues are particular to our work but they also reflect general issues that arise when a project shifts its focus. We offer the following issues because they affect many projects, as they struggle with their identity:

Perceived neglect of the core collection vs. the need to generalize. Many in our classics audience have expressed resentment at our non-classical work and, indeed, the classics collections seem to many to have paid an opportunity cost, as much of our effort has turned elsewhere. In fact, had we not broadened the scope of our approach, we would have even fewer resources to devote to classics, either directly or indirectly

(through general development). Furthermore, the NSDL services grant that is about to begin is based on the automatic linking work that we developed for classics. By stressing the general issues of automatic linking, we have identified a service from classics as a general resource that will serve a vastly larger community. As a result, we are able both to improve the automatic linking service for all users and to embed services widely used by our classicist and humanities users within the much larger NSDL community. This service is now, in our view, more sustainable in the long run.

There is also a general developmental principle at work. Few would argue in principle that we should not share infrastructure, but many of us still resist implementing this in practice. Digital library projects – and especially digital library projects in small, specialized areas of the humanities – must aggressively pursue ways to adapt general solutions that may not at first seem suited to our particular needs [14]. Digital library projects often give up on general solutions too soon, developing their own DTDs, for example, instead of using those of the Text Encoding Initiative, or creating special purpose software, thus solving short-term problems but creating very difficult problems of support over time. We have seen short-term success kill many projects over the past two decades.

Exploring new domains vs. the rigors of disciplinarity. In developing new collections outside of classics, we initially followed our trained instincts and sought leadership from experts in the fields. In working closely with several well-defined expert groups, we found, however, the weight of established practice and tradition to be restrictive. We have shifted instead to seeking advice rather than direction from experts in the field and to offering advice to collection-development projects rather than working under them.

We developed a London collection after making our own assessments and seeking constructive advice. We were able to build a collection that would, we hoped, stimulate ideas about new kinds of intellectual work rather than simply enhance existing research agendas. Such a strategy is risky, since it reflects a vision of what might be useful, and such visions, if they are at all interesting, will often not be productive. Such a strategy is also very hard to fund in a peer reviewed environment, precisely because it follows unconventional paths. The Fund for the Improvement of Postsecondary Education, the IMLS National Leadership Grants, and NSF ITR program are specific attempts to address the challenges of supporting innovative work. The Berger Family Fund for Technology Transfer at Tufts allowed us to establish the London collection.

Our forays into areas such as the history and topography of London, the history of mechanics, early-modern English literature, and others have provoked a range of responses. While interest and excitement are gratifying, the most useful responses have often been the most critical and even hostile. Tangible collections and services provoke concrete discussions based on what does and does not appeal to real users. Digital collections are capital resources that not only retain their value but can evolve over time (e.g., expand, acquire new metadata or tagging). If the base documents are of sufficient interest, data entry is sufficient and document structures solid, publicly accessible collections can attract additional labor and drive debate forward. By intruding into domains beyond classics, we have been able to stimulate thought and debate that would not otherwise have taken to take place.

Using vs. creating digital collections. It is essential that at least some digital library researchers have an opportunity to build and design collections from the ground up. Like the alphabet soup of evaluation forums for language technologies (e.g., TREC, CLEF, ACE, DUC, etc.), most DL research projects work with third-party collections. While prudent, this strategy is sometimes restrictive: researchers can add metadata to documents, but the documents themselves cannot be modified. Because we are interested in how document form and digital library service influence one another, we need to be able to vary the functionality of our collections by modifying every component, including data acquisition, markup, and delivery. Thus while some of our work uses third-party collections, we also have invested substantial labor in creating testbeds of our own, for which we are responsible but which have few if any restrictions. We now have reasonably large, fairly heterogeneous document sets with which we can take risks, without horrifying living authors or even electronic editors of public-domain materials.

The goal of the Perseus Project has been to provide a set of instruments – collections and services – with which we can study new types of use. The remainder of this paper describes some very preliminary findings and points towards the research agenda that will guide us in the final years of our DLI-2–sponsored research. That agenda focuses on answering two basic questions: first, how do digital libraries support various communities as they work now? Second (and far more challenging), how do digital libraries open up new forms of work and, indeed, potentially create new audiences for new ideas?

The latter question is particularly important, because digital libraries have the potential to so alter user expectations that components designed as long-term infrastructure become obsolete. Consider the following scenario: a group of scholars sets to work on a completely new lexicon for a well-studied language, the first in centuries. No one doubts that the existing lexicographic tools would benefit from being replaced. The project manages to acquire five to ten person-years of labor and creates a lexicon, with beautifully composed new entries. The working environment makes good use of existing computational linguistic tools to track collocations and exploit machine-readable versions of the old lexicon. The project delivers its predicted results, and the community of students and scholars working on this language has a far more up-to-date lexicon.

How valuable such a new lexicon would be in five to ten years, however, is not clear. Scholars already read source texts in digital libraries that provide a variety of lexical tools, such as automatically generated links from inflected forms to dictionary entries, morphologically sophisticated searching and analysis tools, and other resources that reduce scholarly dependence upon lexica [15, 16]. We do not know how a new lexicon might aid students learning to read the language, as most student problems center on the relatively simple tasks of finding proper definitions and understanding syntax. Thus while a new lexicon would clearly constitute an advance, the field might have been far better served had it spent its costly time and labor on building a treebank: a database of annotated parses of the most heavily read texts in the corpus [17-19]. A treebank would provide students for the first time with consistent syntactic information about millions of words, potentially a far greater advance than improved lexicon entries. The treebank would also provide a training set for context-free grammars that could be run over tens of millions of words in the remaining corpus, thus providing new research

opportunities. The treebank might even answer more scholarly questions about word usage than general lexicographic articles. If this were so, lexicographers might find themselves becoming not archivists creating fixed textual descriptions, but computational linguists producing dynamic lexical databases. In our view, such a transformation would be salutary, for the servants of inquiry – lexicographers and librarians, traditional and digital – should help inquirers answer the questions we pose now and in the future, rather than serving the needs of an earlier century.

2. AUDIENCES

We have assembled two datasets with which to analyze the needs of the audiences for the collections we have mounted. The Web has clearly allowed new intellectual communities to form, and many intellectual resources formerly available only onsite in special collections are now receiving substantial electronic use (e.g. [20-23]). The 315 million page-accesses we have tracked since 1996 are a useful diachronic dataset that we are beginning to use to track the evolution of user behavior, while the thousands of email messages to webmaster@perseus.tufts.edu trace the reactions of a patron group (albeit a self-selecting one). We have invested substantial staff time in answering as many messages as possible, and this practice has yielded dialogues with a number of users from varying backgrounds.

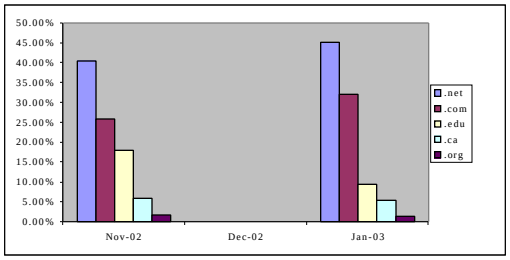


Figure 1: Sample domains for traffic on November 14, 2002 and January 8, 2003.

The bulk of the current Perseus traffic (84%) is concentrated on the established classical collections. This is unsurprising: we have spent fifteen years developing these collections and accompanying services, while developers – including faculty creating syllabi and non-academic web developers – have spent years creating links into the classical digital library. The remaining 16% of the traffic is, however, non-trivial – 1.2 million page accesses in a month of 8 million – and will provide a reasonable basis on which to study some aspects of user behavior.

It is difficult to identify the audience reached by an open site like the Perseus Digital Library. Clearly, those interested in classical antiquity represent a subset of those with a passion for the past. Nevertheless, the audience for Greco-Roman Perseus has consistently surprised us with its sheer size: 6.7 million pages on classical antiquity in a typical month of 8 million aggregate pages. Even more surprising (and gratifying to a classical philologist), approximately 10% of those pages were Greek and Latin source texts in the original, dictionary entries, morphological analyses and other tools that support reading Latin and Greek.

Counting domain extensions is a notoriously problematic instrument for estimating audience composition. Few Perseus hits, for example, come from the .edu domain, but many students access the Internet from accounts with third-party Internet Service Providers, thereby masking the fact that they are pursuing traditional coursework by new electronic means. Nevertheless, the fluctuations of *.edu usage suggests that the number of non-academic users is substantial and, indeed, predominant.

The two dates in Figure 1 reflect very different sections of the academic year. Mid-November is a peak period of student and traditional academic use. Most US universities are, however, on intercession in the first week of January. The percentage of *.edu use declines by only a factor of two. This suggests the possibility that a large percentage of our student traffic comes from third party ISPs and does not show up in the *.edu aggregates. Basic "signal intelligence" may provide a better view of where our traffic lies.

Tracking web usage, we have been able to identify gross patterns: October/November and February/March are the times of peak academic activity. Other patterns emerge as well: note that September 2001 marks the first time when traffic declined for a given month over the previous year. Activity rebounded in October 2001, and it is tempting to hypothesize a "September 11" effect, when academic users were distracted from their expected behavior. Serious system problems emerged in the late summer and fall of 2002. The traffic for October 2002 and 2001 were almost identical. Growth resumed in November and December, as system performance improved, but this improvement lagged beyond what we might otherwise have expected: for the first time, fall totals declined from those of the preceding spring. While the exponential growth in usage clearly cannot continue, this decline probably reflects lingering changes in behavior in response to earlier slow and unreliable system performance.

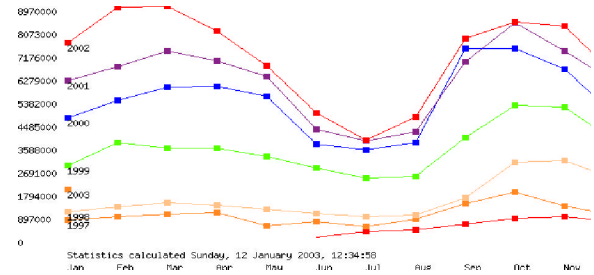


Figure 2: Monthly traffic from mid-1996 through the end of 2002

The variance in usage over the course of the year shows the extent to which the academic calendar influences usage. Traffic varies by a factor of two between peak academic months and July (which is consistently the month of lightest usage). By comparing the number of students enrolled in courses during the summer and those in the traditional academic year, we will probably be able to form a reasonable estimate of the size of our academic audience. Even when this audience is factored out, the non-traditional users will, we expect, remain substantial – far larger than we had expected, given the content and design of the site: even 25% of the July audience would suggest that 1 million pages were sent to individuals beyond academia. Certainly our webmaster correspondence demonstrates the qualitative breadth of the user base. Web log analysis suggests a substantial quantitative base as well.

Counting which documents users read most frequently is an obvious technique. Nevertheless, while the technique may be

obvious, the significance of the results will vary from field to field. If we assume for now that the Perseus Digital Library reaches a representative selection of those reading classical texts, we have the best data ever on who is doing what. Classical texts are capital resources on which scholars have lavished their efforts for centuries. A complex network of commentaries, specialized grammars, and studies exists (although most of this network remains sequestered in print libraries with limited clienteles). We can now see what people are reading and from this begin to understand our audience.

The raw statistics above are only the starting point in a more detailed analysis, which will consider the information needs for each of these audiences. The digital library provides the field of classics with a new instrument by which to see what we are doing and explore what new information resources might be most useful.

3. SERVICES

To provide the services we felt best addressed the needs of our audience, or project found it necessary to create a full-fledged digital library system. Much of our work consisted of adapting general tools to the particular requirements of our collections. Thus, we wanted to be able to perform morphologically aware information retrieval [e.g., search for *fero* (Latin, "to carry") and retrieve *tuli* ("I carried")] with search engines that had no hooks for morphological analysis [24]. Thus we created surrogate files in which we substitute inflected forms for dictionary entries (e.g., *tuli* → *fero*) and search these. When users enter queries, we expand them by substituting the inflected forms for the dictionary

Aeschylus, Agamemnon (aesch. ag.)	5050
Christopher Marlowe, The Tragicall History of D. Faustus (A text) (1999.03.0010)	4635
Euripides, Medea (eur. med.)	4543
Vergil, Aeneid (verg. a. 2)	3549
Sophocles, Antigone (soph. ant.)	3477
Cicero, Against Catiline (cic. catil. catil. 1)	3375

Figure 4: Most commonly viewed primary sources in December 2002. Note the prominence of Marlowe's Faustus (which suggests a class project). The Cicero probably reflects high school Latin readership. Monthly totals for individual primary sources fluctuate far more than for the core reference works.

entries and search for those. The principle was to integrate standard tools into a working system with as little effort as possible.

The rise of robust digital repositories such as FEDORA [25] (which the Tufts library system is adopting) and Dspace [26], the Open Archive Initiative, document-oriented XML searching engines such as HyRex [27-32], and the emerging services component of the National Science Digital Library have, among other things, allowed us to rethink the way that we manage our collections. We have already begun to expose our metadata over the OAI and to integrate OAI services into our own system [33]. We will shift our data object to the Tufts FEDORA repository and translate the services we need into FEDORA. Nevertheless, even as a new generation of digital library infrastructures emerges, there are still services, some particular to cultural heritage digital libraries, that we must still either maintain or develop.

Most of the services described below reflect different forms of information harvesting. We plan to use the OAI to make as many categories of data available as possible: thus, third-party repositories should be able to harvest not only titles and authors but head words from dictionaries and encyclopedias, automatically mined people, places, and dates and any other

Perseus Document (with Perseus id #)	*Hits*
Perseus Encyclopedia (1999.04.0004)	71,784
Charlton T. Lewis, A Latin Dictionary (1999.04.0059)	49,347
Henry George Liddell, A Greek-English Lexicon (1999.04.0057)	49,258
Thomas R. Martin, An Overview of Classical Greek History from Homer to Alexander (1999.04.0009)	44,278
Harry Thurston Peck, Harpers Dictionary of Classical Antiquities (1898) (1999.04.0062)	20,546

Figure 3: Five most commonly viewed texts in December 2002. All five are reference works connected to primary texts by an automatic linking service. Dictionary usage suggests that Perseus users spend as much time reading classical Greek as Latin.

extracted information that third parties would find useful.

The following list enumerates some of the most basic services that we identified. The emerging challenge for digital libraries seems to be multisource, customized summarization: a DL system should be able to determine what supporting information a particular user would require to understand a particular piece of information [31]. Most of the services that we describe below are building blocks for such a system. We have chosen, however, to concentrate on those services that we have either implemented or have funding to develop. We do not emphasize those services already under development in the NSDL but concentrate on issues that particularly characterize the needs of cultural heritage digital libraries.

Document chunking and navigation services. Publishers can pressure authors to follow regular style sheets, thus vastly simplifying the infrastructure required to manage these documents, but cultural heritage documents are structured in a variety of ways, and overlapping hierarchies are common. Thus an edition of Thucydides' *History of the Peloponnesian War* may be structured into traditional book/chapter/section units, suitable for some readers but not for others, who may wish to extract Thucydides' speeches, which can begin and end in the middle of the book/chapter/section units. Some modern scholarly documents have extremely complex structures: the New Variorum Shakespeare series contains two kinds of annotation, a range of narrative essays, and small libraries of source materials, which can themselves be documents of considerable complexity. Providing a reasonable default method of paging through such documents while preserving the flexibility for alternate schemes is often messy and requires systems to look beyond the elegance of BNF formats such as XML.

XML server. At present, most digital libraries tightly couple back-end and front-end, but XML facilitates the separation of content from display. The Perseus Document Management system has for several years worked by extracting well-formed fragments of XML from documents and databases. The task is not always straightforward: extracting lines 21-38 from the scene of a play

can be difficult, since line 21 might appear in the middle of a speech that was itself nested deeply in a complex textual hierarchy. Converting this document management system into an independent XML fragment server will allow third parties to add services and new front-ends to the data that we collect.

Visualization tools. These are crucial but since we and others have published on their importance in previous digital library conferences, we simply allude to these in passing: [21-23, 34-36].

Citation linking. Automatic citation linking has made immense progress [37], but for humanists, locating the document is only a first step. Humanists cite points and (if we are diligent) spans within a document; sometimes these points and spans designate pages or other coarse units, but at other times we use very precise forms of reference (e.g., Homer's Odyssey, Book 9, line 312). In some cases these citations may contain text anchors (Vergil Aen. 1.1: "arma … cano"), which may need to be expanded (e.g, "arma *virumque* cano"). Some works (lexica, for example) may vary the word order or even the words themselves, thus requiring fuzzier matching algorithms. Some citation schemes have remained unchanged for centuries. But others remain in flux (each new edition of Shakespeare, for example, has a tendency to introduce new reference schemes, thus causing tremendous problems for citation systems). Adding to the complexity of automated citation linking, humanists often cite passages using languages or spelling conventions different from those in the original sources. At the very least, software supporting pre-nineteenth century scholarship must be able to match original spelling against modernized forms.

Quotation identification and source tracking (not implemented in Perseus). A corollary to the citation-linking problem is that of quotation identification. Many documents in the humanities contain quotations from earlier sources: almost half of the initial ten million words in the London collection was quoted text. Most of these quotes have no precise references; many do not even mention precise authors. A DL system needs to be able to search quickly and automatically for likely sources for quotations. Such a system should also scan for unquoted sources (for which plagiarism detection services provide a technical model [38]).

Named entity identification and analysis of "encyclopedic data." People, places, and things may have bored history students for millennia, but they remain key components of historical documents. Perseus scans for dates and place names in all full text documents [36, 39, 40]. The service at present is limited to English (in part because we have English translations for the vast majority of source texts), but it enables us to provide automatically generated (and hence scalable) timelines and maps to help users assess the contents of collections, detect events, [41], [42], [34, 35] and ultimately search by time and place (e.g., "documents relevant to Worcester County, Mass. in the 1840s"). We have also begun to track personal names, monetary sums, and other readily identifiable entities.

Named-entity identification is notoriously domain specific [43-45]. Even such apparently universal entities as people, places, and things can vary widely from culture to culture. Western audiences in 2001, for example, learned that Afghans often have only a single name. Greco-Roman texts do not, of course, use modern year numbers or even easily followed month/day schemes. Digital library systems need to be able to apply different information extraction routines to different collections.

We have begun to see work on integrating information extraction into digital libraries (e.g. [46-48]), but this task has barely begun. Much named-entity work depends upon heuristics written in application-specific formats: very little work has gone into making such core information portable, much less in creating sharable digital libraries of named-entity heuristics. Just as cultural-heritage scholars have traditionally spent years preparing editions for publication, so corpus editors will spend years developing gazetteers, heuristics, and training sets for large corpora, and they will need be able to exchange and build on one another's efforts over many years [49, 50].

Semantic services. A cultural heritage digital library system should automatically integrate new texts, lexica, grammars, treebanks, and other resources into linguistic services which are constantly updated. We have implemented such services for collocations of Greek and Latin words [51, 52]; a mature DL system would harvest new documents to enhance cross-language information retrieval, machine translation, automatic summarization, and other services.

Authority-list editors (not yet implemented but under development as an NSDL service under [13]): Information-extraction systems can recognize that "Mark Twain" and "Twain" in close proximity are both personal names and probably refer to the same person. But we need tools that connect both references to a general authority list, one which can also help information-extraction systems recognize that "Mark Twain" and "Samuel Clemens" are both instantiations of "Twain, Mark, 1835-1910." Likewise, on a semantic level, we need to be able to express the fact that "bank" in document A is a financial institution, not the edge of a river, and to connect this instance to a definition in a third-party dictionary. Such an authority-list editor should be able to predict the most likely meaning in the case of multiple instances.

Runtime automatic linking. Some texts will have complex pre-established markup associated with them, but we also need to be able to identify and add informative links to key words and phrases on the fly. Such automatic linking has been part of the Perseus system for more than ten years [24, 53-57], but this feature will be expanded and refined to become a service within the NSDL under [13].

Automatic evaluation services. Evaluating digital libraries in general [58], and language technology services in particular, is difficult [29, 59-62], but because digital libraries contain many knowledge resources, such as encyclopedias, indices, and lexica, that include manually verified data, digital library systems should be able to mine their manual resources to refine their automated services.

Automated comparisons between information extraction and manual indices reveal interesting differences between human practice and machine performance: humans are better at interpreting indirect references — "the South's best hope" as a reference to Robert E. Lee, for example — but our information extraction services had much better recall than the manual indices. The precision, recall and F-measure numbers listed in Figure 5 thus generally underestimate the performance of information extraction routines. Nevertheless, while the numbers below may be questionable as absolute measures, the automated measures seem likely to gauge performance between comparable information extraction systems. Much work needs to be done in

developing DL services that mine such pre-existing manual data to track new service performance.

Work	*Precision*	*Recall*	*F-Measure*
blcw01	0.877	0.904	0.890
blcw02	0.885	0.908	0.896
blcw03	0.871	0.872	0.871
blcw04	0.894	0.878	0.886
phcw01	0.812	0.937	0.870
phcw02	0.578	0.941	0.716
phcw03	0.609	0.940	0.739
phcw04	0.660	0.895	0.760
phcw05	0.591	0.944	0.727
phcw06	0.801	0.943	0.866
phcw07	0.789	0.935	0.856
phcw08	0.795	0.945	0.864
phcw09	0.848	0.931	0.888
phcw10	0.982	0.949	0.965
rebrec.diary	0.698	0.929	0.797

Figure 5: Precision, recall and F-measures (calculated here as 2 * precision * recall / precision + recall) for identification of personal names in a series of Civil War books: blcw = *Battles and Leaders of the Civil War* [63]; the *Photographic History of the Civil War* [64]; and the diary sections of the multivolume *Rebellion Record* [65]. The DL system automatically calculates these by comparing the output of the named-entity tagger with the contents of on-line indices.

4. COLLECTIONS

The Perseus Digital Library comprises third-party collections and those we have created for experimental purposes. While the Perseus Digital Library includes non-textual materials that are both well understood (color images and their accompanying metadata) and experimental (collections of GIS and 3D materials), we focus here on texts. Where we have the rights to do so, we will make these collections accessible as testbeds to those conducting research in language technologies and digital libraries. All Perseus data is scheduled to become part of the Tufts University FEDORA repository [25], which will provide a long-term source on which third party researchers and developers can rely.

At the moment, we have seven substantial collections. Two of these are from third parties; the other five were developed partly or entirely by Perseus.

Each of these collections allows us to experiment with a different domain, a different cluster of audiences, and a different set of research challenges. The London collection, for example, allows us to study problems and opportunities of a geospatially oriented DL with a small geographic focus and centuries of activity. The US Civil War collection, by contrast, is geographically dispersed but temporally compact. Both collections have substantial popular audiences and thus offer greater opportunities for outreach than the Greco-Roman materials. The Early Modern English collection taps into a scholarly infrastructure that superficially resembles that developed for classics, but students of early modern culture in general and Shakespeare in particular have expectations that differ from each other and from classics. Where much of our efforts have focused on extracting people, places, dates and other encyclopedic data, the history of mechanics collection has a very different orientation, forcing us to consider how to track and analyze formulas, technical language and broader mental models. The American Memory collections allow us to compare our work with mainstream US collection development, while the Duke Databank of Documentary Papyri is a core resource for an intense, highly organized subdiscipline of classical research.

The eclectic set of collections thus forces us to confront a wide range of challenges, not the least of which is the management of heterogeneous materials. We built the Perseus Digital Library precisely to explore these difficulties, and the major theme of our future work will be designing services that work with diverse collections and audiences.

Elsewhere we have discussed the process of boot-strapping digital collections: to achieve integrated systems in which documents interact not only with users but with each other [66], we place particular emphasis on dictionaries, encyclopedias, handbooks, and other reference materials [39, 40, 67-71]. While we can work with documents in a variety of formats (including free text, HTML, PDF, and RTF), we continue to explore how structured markup and digital library services co-evolve. Our services, current and envisioned, shape the structure of documents, while document structure enables services. Every markup tag is both an interpretation and an investment. In the digital library context, each tag represents a statement by the collection designer that some digital library service either exists or will exist to exploit it. By creating documents and services we have been able to see in many concrete ways how the two interact.

Collection	*Source*	*Size (million words)*
Classics	Perseus	50
US Civil War and other 19th Century Materials	Perseus	41
Selected American Memory Collections	Third party	38
Archimedes/History of Mechanics	Perseus*	15
London Collection	Perseus*	13
Early Modern English	Perseus	7.3
Duke Databank of Documentary Papyri	Third party	4

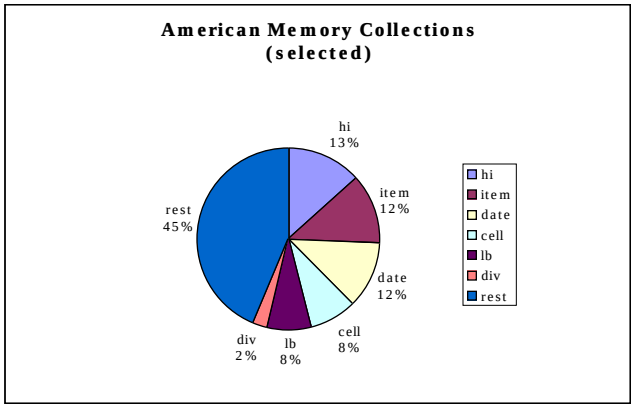
Figure 6: Perseus's major collections. An asterisk (*) denotes a collection developed jointly by Perseus and a third party.

Consider the markup of quotations, for example. Most projects leave quotations unmarked, but our use of the TEI `Quote` tag forces us to make certain that each quotation is well-formed. Because quotation marks are a major source of typographical error in both transcriptions and original sources, ensuring well-formedness, even with the assistance of reasonable software tools, is labor-intensive and thus expensive. But identifiable quotations are a valuable asset, for they can be used to discover or verify citations and cross-references, both explicit and implicit, across multiple collections. The `Quote` tag thus represents an

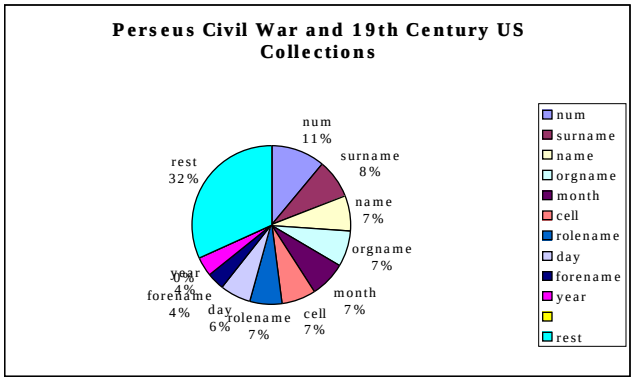
investment for a service that we have not yet implemented but can easily envision.[1]

We have devoted substantial effort to developing automatic markup engines to support digital library services. The following tables describe the density and frequency of markup in four very different collections, and provide some sense of how TEI conformant collections in a single digital library may differ.

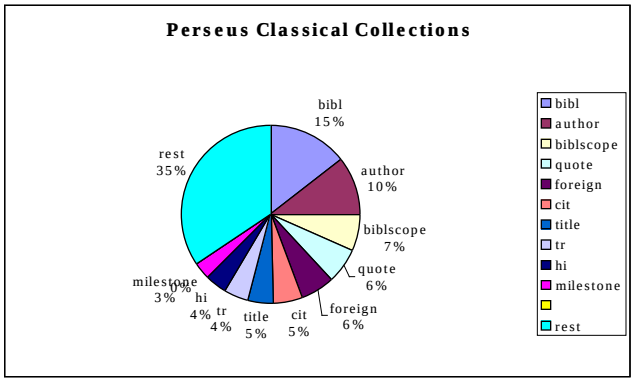
Counting tags is, of course, problematic. Markup quickly clogs files, and information extraction systems have inspired work on more tractable standoff markup schemes [47, 48, 75-78]. Ten years ago, when the first classical Perseus CD ROMs were produced, we were already using draft TEI guidelines but were generating at least one linguistic tag of standoff markup for each of the million words of Greek [54, 79]. Today we scan for possible place names in all collections that have no explicit geographic tags and store these in standoff markup: the number of tags associated with the American Memory texts is thus already much larger than those encoded in the files. Nevertheless, the tables below provide overviews of how each collection differs and what sorts of services/audiences each is aimed at. Note that all three Perseus collections thus have roughly five times as many tags per word as the American Memory collections. Note, however, that the automatically generated tags are not manually checked and thus contain many inaccuracies [36].



American Memory Collections on California, Chesapeake Bay, and the Upper Midwest. These collections provide textbook examples of well-executed level 4 TEI markup, which focuses on encoding the structural elements of a text [80]. The HI tag simply encodes that the source text was in italics and puts off the task of determining where the italics indicates book titles, foreign language quotes, etc. The `Cell` tags indicate parts of tables and reflect page layout. Ultimately, tables need to be analyzed, since most table rows are, in fact, nascent database records. Note, however, the large number of `Date` tags – these tagged dates allow us to determine the temporal coverage of individual documents within the American Memory collections and are an extremely valuable addition. They represent a judicious cost/benefit judgment, expanding judiciously beyond level 4 markup.

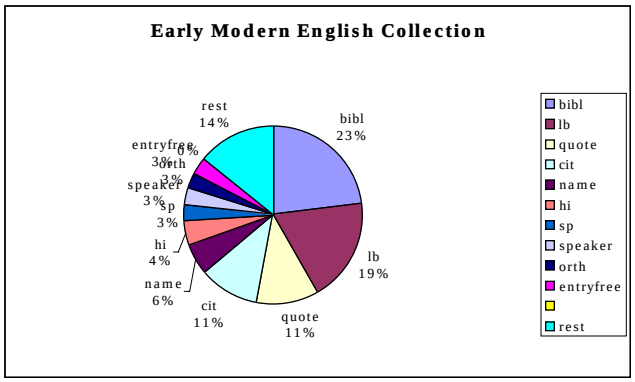


Perseus Civil War and Nineteenth Century US Collections. All of the ten most common tags are automatically generated. The `Num` tag simply brackets every floating number (though the tagger does do a good job of mapping "21" and "twenty-one" etc. to the same quantity). Bracketing tags such as `DateStruct`, `PersName` etc. have been ignored so as not to count tags twice. The personal name and date routines were fairly general. Most of the `OrgName` tags represent regiments and other military units and required domain specific routines. Most of the `Name` tags are qualified with the attribute `Type=place`. Thus, we have located more than 1.5 million possible surnames, places and organizations in the 41 million-word collection.



Perseus Classical Collections. Two monumental research lexica form core resources for the classical collections. The 1.1 million `Bibl` tags represent more than a million point-to-point links. The 400,000 `Cit` tags, along with 100,000 annotation tags (not numerous enough to make the top 10) indicate that we have 500,000 passages where we can associate particular text spans and citations (thus providing data for the 3 dimensional citation links described above). The 250,000 `Milestone` tags indicate anchors for citation/chunking schemes that lay outside the dominant XML hierarchies of the texts. While the density of tags is high (15 tags for every 100 words), the classical texts have less automatic internal markup than the Civil War collection. Hence the semantically vacuous `Hi` tag, an artifact of the data-entry process signaling text that remains to be disambiguated, is in the top ten.

[1] It also constitutes a bet that the cost of well-formed quotations will be less than that of correcting for the problems that arise when we try to extract quotations from texts with unverified quotation marks. Some of our recent research suggests we may lose this bet.



Early Modern English Collection. The early modern collection is particularly heterogeneous, including the plays of Marlowe and Shakespeare, as well as substantial prose works such as Holinshed's *Chronicles* and Hakluyt's *Voyages*. Most of the tags in the early modern texts were added either by the data entry firm or by semi-automated post-processing. Nevertheless, the density of tags is high, reflecting the labor costs of production. The Shakespeare community is accustomed to highly finished, manually produced reference works and has proven the most demanding of all of our audiences. The `Sp` and `Speaker` tags reflect the prominence of dramatic works in the corpus. Three on-line Shakespeare lexica also account for most of the top ten tags.

5. CONCLUSIONS

We have provided an overview of the three core topics of our research: how networked digital libraries serve established but also reach new audiences; the services that cultural heritage digital libraries need to support; and the document structures that provide the basis for these services. Like the NSDL, cultural heritage digital libraries have a broad, "K to gray" audience. But cultural heritage digital libraries play a particularly important role for students of the humanities, because the digital library is a primary laboratory and space for research. The sources within a cultural heritage digital library constitute primary data. Reading support, including both automatic linguistic services for multilingual and summarization services for multilingual and monolingual reading, is especially important. The clear focus on reading support has allowed us to begin contributing to the NSDL as well [13]. Because cultural heritage collections do not become obsolete but become, if anything, more valuable evidence as the past recedes, humanists have spent generations (and, in some cases, millennia) creating paper-based knowledge bases on some topics. We have capitalized on this phenomenon, creating digital collections where we could mine rich sets of paper sources and begin establishing connections with on-going communities. The resulting system has begun to bring together diverse audiences, services and collections. The resulting interactions are both an object for analysis and an instrument for new collaborations.

6. ACKNOWLEDGMENTS

A grant from the Digital Library Initiative Phase 2 (NSF IIS-9817484), with particular backing in our case from the National Endowment for the Humanities, provided the primary support for this work.

7. REFERENCES

1. Palmer, C.L. and L. Neumann, *The Reseach Work of Interdisciplinary Humanities Scholars: Exploration and Translation.* Library Quarterly, 2002. **72**(1): p. 85-117
2. Brockman, W.S., et al., *Scholarly work in the humanities and the evolving information environment.* 2001, Digital Library Federation and the Council on Library and Information Resources: Washington, DC.
3. Tibbo, H.R. *Primarily History: Historians and the Search for Primary Source Materials*. in *JCDL 2002: Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries*. 2002. Portland, OR http://doi.acm.org/10.1145/544220.544262.
4. Marshall, C.C. and C. Ruotolo. *Reading-in-the-small: a study of reading on small form factor devices.* in *JCDL 2002: Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries*. 2002. Portland, OR http://doi.acm.org/10.1145/544220.544262.
5. Wolfe, J.L. *Effects of annotations on student readers and writers.* in *Proceedings of the eleventh ACM conference on Hypertext and hypermedia*. 2000. San Antonio, TX USA: ACM Press http://www.acm.org/pubs/articles/proceedings/hypertext/336296/p143-modha/p143-modha.pdf.
6. Rosenzweig, R., *Everyone a Historian -- afterthoughts to the Presence of the Past.* 2002, George Mason University.
7. Toplin, R.B., *Ken Burns's The Civil War : the historian's response*. 1996, New York: Oxford University Press. xxvii, 197.
8. Rosenzweig, R. and D.P. Thelen, *The presence of the past : popular uses of history in American life -- supplementary web site*. 1998, George Mason University: Fairfax, VA.
9. Rosenzweig, R. and D.P. Thelen, *The presence of the past : popular uses of history in American life*. 1998, New York: Columbia University Press. x, 291.
10. Burns, K., et al., *The Civil War*. 1989, PBS Video: Alexandria, VA.
11. Aucoin, D., *Fast Forward into the Past: What do we learn when TV becomes our history teacher?*, in *Boston Globe*. 1997: Boston, MA. p. P1, P4.
12. Crane, G., *Thucydides and the Ancient Simplicity*. 1997, Berkeley and Los Angeles: University of California Press.
13. Colati, G., G. Crane, and S. Choudhury, *Managing Authority Lists for Customized Linking and Visualization: a Service for the National STEM Digital Library*. 2002.
14. Mahoney, A., et al. *Generalizing the Perseus XML document manager*. in *Linguistic exploration: workshop on web-based language documentation and description*. 2000. Philadelphia http://www.ldc.upenn.edu/exploration/expl2000/papers/mahoney/mahoney.htm.
15. Crane, G., *New Technologies for Reading: the Lexicon and the Digital Library.* Classical World, 1998. **92**: p. 471-501
16. Smith, D., J.A. Rydberg-Cox, and G. Crane, *The Perseus Project: A Digital Library for the Humanities.* Literary and Linguistic Computing, 2000. **15**(1): p. 15-25
17. Marcus, M., et al. *The Penn TREEBANK: Annotating predicate argument structure*. in *ARPA '94*. 1994 http://www.ldc.upenn.edu/doc/treebank2/arpa94.html.
18. Marcus, M.P., B. Santorini, and M.A. Marcinkiewicz, *Building a large annotated corpus of English: the Penn Treebank.* Computational Linguistics, 1993. **19**: p. 22 http://www.ldc.upenn.edu/Catalog/docs/treebank2/cl93.html.
19. Chiou, F.-D., D. Chiang, and M. Palmer. *Facilitating Treebank Annotation Using a Statistical Parser*. in *HLT2001*. 2001. San Diego http://www.hlt2001.org/papers/hlt2001-26.pdf.
20. Choudhury, G.S., *Strike up the score: deriving searchable and playable digital formats from sheet music.* D-Lib Magazine, 2001. **7**(2) http://www.dlib.org/dlib/february01/choudhury/02choudhury.html.
21. Kochumman, R., et al. *Towards an Electronic Variorum Edition of Cervantes' Don Quixote: Visualizations that support preparation.* in *JCDL 2002: Proceedings of the second ACM/IEEE-CS joint*

conference on Digital libraries. 2002. Portland, OR http://doi.acm.org/10.1145/544220.544262.

22. Monroy, C., et al. *Visualization of variants in textual collations to analyze the evolution of literary works in the Cervantes project*. in *6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2002)*. 2002. Rome

23. 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

24. Crane, G., *Generating and Parsing Classical Greek.* Literary and Linguistic Computing, 1991. **6**: p. 243-245

25. Payette, S. and T. Staples. *The Mellon Fedora Project: Digital LIbrary Architecture Meets XML and Web Services*. in *European Conference on Digital Libraries (ECDL)*. 2002. Rome: Springer

26. Smith, M. *DSpace: an institutional repository from the MIT Libraries and Hewlett Packard Laboratories*. in *European Conference on Digital Libraries (ECDL)*. 2002. Rome: Springer

27. Fuhr, N. and K. Großjohann, *XIRQL: A Query Language for Information Retrieval in XML*, in *Proceedings of the 24th Annual International Conference on Research and development in Information Retrieval*, B. Croft, et al., Editors. 2001, ACM: New York. p. 172-180.

28. Abolhassani, M., et al., *HyREX: Hypermedia Retrieval Engine for XML*. 2002, University of Dortmund: Dortmund.

29. Fuhr, N., M. Lalmas, and G. Kazai, *INEX: Initiative for the Evaluation of XML retrieval*. 2002, University of Dortmund.

30. Fuhr, N. and G. Weikum, *Classification and Intelligent Search on Information in XML.* IEEE Data Engineering Bulletin, 2002. **25**(1): p. 51-58 http://ls6-www.informatik.uni-dortmund.de/bib/fulltext/ir/Fuhr_Weikum:02.pdf.

31. Fuhr, N., et al. *Daffodil: An Integrated Desktop for Supporting High-Level Search Activities in Federated Digital Libraries.* in *European Conference on Digital Libraries (ECDL)*. 2002. Rome: Springer

32. Fuhr, N., N. Gövert, and K. Großjohann. *HyREX: Hyper-media Retrieval Engine for XML*. in *SIGIR 2002*. 2002. Tampere, Finland: ACM http://patty.informatik.uni-duisburg.de/bib/fulltext/ir/Fuhr_etal:02b.pdf.

33. Smith, D.A., A. Mahoney, and G. Crane. *Integrating Harvesting into Digital Library Content*. in *JCDL 2002: Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries*. 2002. Portland, OR http://doi.acm.org/10.1145/544220.544262.

34. Smith, D.A. *Detecting events with date and place information in unstructured text*. in *2nd ACM+IEEE Joint Conference on Digital Libraries*. 2002. Portland, OR: ACM Press http://www.perseus.tufts.edu/Articles/datestat.pdf.

35. Smith, D.A. *Detecting and browsing events in unstructured text*. in *Proceedings of the 25th Annual ACM SIGIR Conference*. 2002. Tampere, Finland: ACM http://www.perseus.tufts.edu/Articles/eventir.pdf.

36. Smith, D.A. and G.R. Crane, *Disambiguating Geographic Names in a Historical Digital Library*. 2001, Perseus Project/Tufts University: Medford, MA.

37. Hitchcock, S., et al., *Open Citation Linking: the way forward.* D-Lib Magazine, 2002. **8**(10) http://www.dlib.org/dlib/october02/hitchcock/10hitchcock.html.

38. 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

39. 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

40. Crane, G., et al., *Drudgery and Deep Thought: Designing Digital Libraries for the Humanities.* Communications of the ACM, 2001. **44**(5)

41. Larson, R.R., *Geographic Information Retrieval and Spatial Browsing*, in *Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information*, L.C.S.a.M. Gluck, Editor. {April} 1995. p. 81-123.

42. Baldonado, M.Q.W. and T. Winograd. *Hi-cites: dynamically created citations with active highlighting*. in *Conference proceedings on Human factors in computing systems*. 1998: ACM Press http://www.acm.org/pubs/articles/proceedings/chi/274644/p408-baldonado/p408-baldonado.pdf.

43. Hattunen, S. and Satoshi. *Diversity of scenarios in information extraction*. in *LREC 2002: Third International Conference on Language Resources and Evaluation*. 2002. Las Palmas, Canary Islands, Spain http://nlp.cs.nyu.edu/publication/papers/huttunen-lrec02.pdf.

44. Hirschman, L., et al. *Integrated Feasibility Experiment for Bio-Security: IFE-Bio, A TIDES Demonstration*. in *HLT2001*. 2001. San Diego, CA http://www.hlt2001.org/papers/hlt2001-38.pdf.

45. Gaizauskas, R. *Intelligent access to text: integrating information extraction technology into text browsers.* in *HLT2001*. 2001. San Diego, CA http://www.hlt2001.org/papers/hlt2001-36.pdf.

46. Bontcheva, K., et al. *Using Human Language Technology for Automatic Annotation and Indexing of Digital Library Content.* in *6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2002)*. 2002. Rome http://gate.ac.uk/sale/ecdl02/ecdl.pdf.

47. Cunningham, H., et al., *Developing Language Processing Components with GATE (a User Guide)*. 2002, The University of Sheffield: Sheffield, UK.

48. Maynard, D., et al. *Adapting a robust multi-genre NE system for automatic content extraction.* in *Tenth International Conference on Artificial Intelligence: Methodology, Systems, Applications*. 2002 http://gate.ac.uk/sale/talks/aimsa02/01.html.

49. 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

50. 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

51. Rydberg-Cox, J.A., *Word Co-Occurrence and Lexical Acquisition in Ancient Greek Texts.* Literary and Linguistic Computing, 2000. **15**(2): p. 121-129

52. Smith, D.A., A. Mahoney, and J.A. Rydberg-Cox, *Management of XML Documents in an Integrated Digital Library*, in *Proceedings of Extreme Markup Languages 2000*. 2000: Montreal. p. 219-224.

53. Crane, G., *Redefining the Book: Some Preliminary Problems.* Academic Computing, 1988(February)

54. Crane, G., ed. *Perseus 1.0: Interactive Sources and Studies on Ancient Greek Culture*. 1992, Yale University Press: New Haven, CT.

55. Crane, G., *What is Perseus? What is it not? Comments on the Bryn Mawr Review of Perseus 1.0.* BMCR, 1992. **3**(6): p. 497-502

56. Crane, G., *The Perseus Project: An Evolving Digital Library*. 2000, Tufts University.

57. Crane, G. *Building a Digital Library: the Perseus Project as a Case Study in the Humanities*. in *Proceedings of the 1st ACM International Conference on Digital Libraries*. 1996: ACM

58. Marchionini, G., *Evaluting Digital Libraries: A Longitudinal and Multi-facted View*. Library Trends, 2001. **49**(2): p. 304-333

59. Voorhees, E.M. *Overview of TREC 2001*. in *TREC 2001*. 2001. Gaithersburg, MD 20899: NIST http://trec.nist.gov/pubs/trec10/papers/overview_10.pdf.

60. *ACE Evaluation plan version 06*. 2002.

61. Hovy, E., M. King, and A. Popescu-Belis. *An introduction to MT evaluation*. in *Machine translation evaluation workshop: LREC 2002: Third International conference on language resources and evaluation*. 2002. Las Palmas, Canary Islands http://www.issco.unige.ch/projects/isle/mteval-may02/mteval-lrec2002.pdf.

62. Dabbadie, M., et al. *A hands-on study of the reliability and coherence of evaluation metrics*. in *Machine translation evaluation workshop: LREC 2002: Third International conference on language resources and evaluation*. 2002. Las Palmas, Canary Islands http://www.issco.unige.ch/projects/isle/mteval-may02/mteval-lrec2002.pdf.

63. Johnson, R.U., et al., *Battles and leaders of the Civil War ; being for the most part contributions by Union and Confederate officers : based upon "The Century war series" edited by Robert Underwood Johnson and Clarence Clough Buel*. 1887, New-York: Century Co. 32 v. in 4.

64. Miller, F.T. and R.S. Lanier, *The photographic history of the civil war*. 1911, New York: Review of Reviews. 10 v.

65. Moore, F. and McLellan Lincoln Collection (Brown University), *The Rebellion record; a diary of American events*. 1861, New York,: G. P. Putnam D. Van Nostrand. 11 v.

66. Soergel, D., *A framework for digital library research: broadening the vision*. D-Lib Magazine, 2002. **8**(12) http://www.dlib.org/dlib/december02/soergel/12soergel.html.

67. Crane, G. *Cultural Heritage Digital Libraries: Needs and Components*. in *European Conference on Digital Libraries*. 2002. Rome: Springer

68. Crane, G., *In a digital world, no book is an island: designing electronic primary sources and reference works for the humanities*, in *Creation, Use and Deployment of Digital Information*, L. Breure and A. Dillon, Editors. 2002, Lawrence Earlbaum Associates. p. forthcoming.

69. Crane, G., *Designing Documents to Enhance the Performance of Digital Libraries: Time, Space, People and a Digital Library of London*. D-Lib Magazine, 2000. **6**(7/8)

70. Crane, G., *Extending a Digital Library: Beginning a Roman Perseus*. New England Classical Journal, 2000. **27**(3): p. 140-160 http://www.perseus.tufts.edu/cgi-bin/ptext?doc=2000.06.0003.

71. Crane, G., et al., *The symbiosis between content and technology in the Perseus Digital Library*. Cultivate Interactive, 2000. **1**(2) http://www.cultivate-int.org/issue2/perseus/.

72. Sperberg-McQueen, C.M. and L. Burnard, eds. *Guidelines for Electronic Text Encoding and Interchange*. 1994, Text Encoding Initiative: Chicago and Oxford.

73. Sperberg-McQueen, C.M. and L. Burnard, eds. *TEI P4: Guidelines for Electronic Text Encoding and Interchange -- XML-compatible version*. 2001, TEI-Consortium.

74. Sperberg-McQueen, C.M. and L. Burnard, *Guidelines for electronic text encoding and interchange*. 1994, Electronic Book Technologies: Providence, RI.

75. Anand, P., et al. *Qanda and the Catalyst Architecture*. in *The Tenth Text REtrieval Conference (TREC 2001)*. 2001. Gaithersburg, MD: Department of Commerce, National Institute of Standards and Technology http://trec.nist.gov/pubs/trec10/papers/MITRE-trecX-1.pdf.

76. Bird, S. and M. Liberman, *A formal framework for linguistic annotation*. Speech Communication, 2001. **33**(1,2): p. 23-60 http://arXiv.org/abs/cs/0010033.

77. Bird, S., et al. *TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse Tools Built on the Annotation Graph Toolkit*. in *Proceedings of the Third International Conference on Language Resources and Evaluation, European Language Resources Association*. 2002. Paris http://arXiv.org/abs/cs.CL/0204006.

78. Cotton, S. and S. Bird. *An Integrated Framework for Treebanks and Multilayer Annotations*. in *Third International Conference on Language Resources and Evaluation. European Language Resources Association*. 2002. Paris http://arXiv.org/abs/cs.CL/0204007.

79. Mylonas, E., et al., *The Perseus Project: Data in the Electronic Age*, in *Computing and the Classics*. 1991, University of Arizona Press: Tucson. p. forthcoming.

80. Friedland, L., et al., *TEI Text Encoding in Libraries: Draft Guidelines for Best Encoding Practices (Version 1.0)*. 1999.