Report on the Delos/NSF working group on Emerging Language Technologies and the Rediscovery of the Past: A Research Agenda

Authors:
Gregory Crane, Tufts University, USA,
Kalina Bontcheva, University of Sheffield, UK

EU Participants:
Jeremy Black, Oxford, UK
Andera Bozzi, CNR/Pisa, Italy
Brian Fuchs, Max Planck Institute for the History of Science, Berlin, Germany
Kurt Gaertner, University of Trier, Germany
Susan Hockey, University College London, UK
Dolores Iorizzo, Imperial College, London, UK
Rüdiger Niehl, Heidelberg, Germany
Stefan Rueger, Imperial College, London, UK

US Participants:
Jason, Eisner, Johns Hopkins
Jay Ponte, Mitre Corporation
Jeffrey Rydberg-Cox, University of Missouri at Kansas City
Peter Scharf, Brown University
David Smith, Johns Hopkins
Clifford Wulfman, Tutfts University

Summary

Our Delos/NSF working group explored the possibilities that emerging language technologies open up for teaching, learning and research in the broad area of cultural heritage. While students of the past stand to gain powerful new tools that will affect every aspect of their work, developers of language technology also would benefit from exploring the needs of new audiences and new collections. While multi-lingual technologies may prove the most revolutionary, this initial report focuses on mono-lingual technologies such as information extraction, summarization and other aspects of document of understanding. We describe some of the audiences affected and list technologies for evaluation.

Introduction

Researchers from a range of disciplines have spent decades studying how computational technologies can be used to study, analyze, process, and understand human language and its products, speech and text. Single-document summarization, multi-document clustering, cross-language information retrieval, named-entity identification and tracking, question-answering: all these tasks have been the focus of intensive research by linguists, computer scientists, statisticians, acousticians, and others. A veritable alphabet soup of forums has arisen to evaluate solutions to these tasks, among which are TREC (Text Retrieval Conference), ACE (Automatic Content Extraction), DUC (Document Understanding Conference, which succeeded MUC: Message Understanding Conference), and CLEF (Cross Language Evaluation Forum).

Out of these evaluation forums has come a set of test collections, which are used to evaluate specific technologies. Because much, if not most, of the funding for all this research comes from DARPA, through its TIDES project, much of the research has focused on the problems of intelligence analysis; most of these test

collections are derived from modern news sources, and comprise relatively brief texts, written by professional journalists, about recent events.[1] Relatively little work has gone into studying how these "language technologies" might be applied to the oceans of historical source material in our museums, libraries, and other repositories large and small.

The thirst for these materials is prodigious. Students in primary, secondary, and postsecondary school constitute a large audience for history, but much of the interest is extracurricular. In a recent survey of 1,500 Americans, two fifths spoke passionately about hobbies, collections, or other pursuits related to the past [1, 2]. Broadcasters like PBS and the History Channel produce a steady stream of new programming on historical topics. The National Endowment of the Humanities, the Institute of Museum and Library Services, the Library of Congress, the National Park Service, and other federal agencies are all supporting the production of new digital resources about the past. Historical societies include not only well-established professional organizations such as the Massachusetts Historical Society but also hundreds, if not thousands, of small local groups run by volunteers [3]. Many state and city governments have offices that oversee and provide information about historic districts and structures. [2] We believe it is safe to say that Americans have an insatiable craving for history.

Apart from the general public, present day scholars, e.g., historians of science, also need new ways to access and annotate the large digital collections that are now becoming available, e.g., the Newton digital library. Existing searching technology, based on Information Retrieval (IR), is too limiting for certain types of queries, e.g., "find all kings and queens mentioned in the works of Newton and show the relevant paragraphs in the documents". An IR search engine cannot deal with this query because it is not based on keywords, i.e., just searching for "king & queen" will not be enough, because these words do not always appear in the text preceding their names. However, language technologies, such as entity detection and tracking, can help scholars to answer such queries, thus allowing them more powerful ways of studying their data.

This report, then, offers recommendations on how best to deploy language technologies to tap the vast reserves of historical source materials and bring the fruits of decades of costly research back to a public with an insatiable craving for history.

We recommend concentrating on three core groups from within the broad constituencies listed above: producers of informational resources, developers of language technologies that will help make those resources available to wider audiences, and those end users themselves. Our hypothesis is that all three groups can learn from each other. Data and service providers clearly cannot know too much about the needs of their audiences, but more important, perhaps, is an expanded dialogue between developers of historical content and those developing the language technologies by means of which users will increasingly experience that content.

Users: The user communities described above are large and their needs are diverse. We recommend concentrating on the following questions:

- What are the current needs of these communities? Where do needs substantively diverge and where are there hidden commonalities of interest?
- What uses will emerge as new infrastructures (e.g., increasing wireless connectivity and more sophisticated hand-held devices) take shape?

---

[1] Substantial work is also being performed in the medical field as well: for example, the GENIA Project (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/) PIR: Protein Information Resource (http://pir.georgetown.edu/pirwww/), the TRESTLE Project (http://nlp.shef.ac.uk/trestle/), and various parts of the work on medical informatics at Stanford (http://www.smi.stanford.edu/).

[2] Private individuals and businesses own most of the sites listed in the national historic register, but few have ready access to the information about their own properties. For this audience we can consider not only intellectual curiosity but economic impact – many historic districts owe their existence to business investment and a conviction that the history can have economic value [4-7].

New technology needs to serve current, clearly perceived needs if it is to be widely adopted, but language technologies have the potential to stimulate entirely new areas of inquiry. We must seek to identify methods, current and potential, in which information science research on language technologies transforms historical inquiry throughout society.

While many different communities need to be studied, we recommend beginning by examining two variables: (1) background expertise and (2) attitude to the task at hand. Sixth graders preparing reports on the anti-slavery movement have different needs from professional researchers, but professional academics engaged in interdisciplinary research also need basic information and quick summaries. Nor are professional researchers necessarily the most dedicated, knowledgeable and demanding users: students of the particular historical periods,

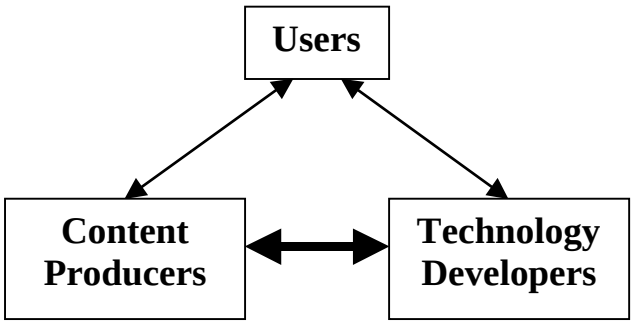

Figure 1: Users interact both with content and the technological infrastructure by which they access that content. This report considers not only the user/content and user/technology relationships but stimulates a third conversation between the content producers and the technology developers. This project will promote new collaborations between content producers and technology developers, as these communities understand each other's needs and possibilities.

members of local historical and genealogical societies, and other amateurs often ask very different questions from their professional colleagues, but their knowledge is often deep and their energies immense. Academics may, for example, bemoan the fascination with the military history of the Civil War (see, for example, the reaction of academic historians to Ken Burns' *Civil War* series [8, 9]), but many acknowledge the expertise amateurs have laboriously acquired. Popular historical interests may not follow conventional academic tracks, but they are widespread and vigorously pursued.[1, 10].

Developers of Language Technologies: Language technologists know all too well that many of their techniques need to be tuned for varying domains – reducing the labor required in shifting from one domain to another is, in fact, a major focus of current research [11-13]. It is vital that we analyze the similarities and differences among the materials of different cultures and domains and identify the needs of a diverse range of users working with these materials. Historical documents vary widely in genre, size, and, indeed, every feature, since the historical records potentially includes all prior published materials that currently survive. Cultural heritage materials include corpora that predate numerical street addresses, clock time, precise calendar dates, and other named entities, so specific challenges – event identification, geographical disambiguation, named-entity resolution – vary in complexity from domain to domain.

Content Producers: Technology inspires the production, and constrains the form, of content. The Web spawned a dizzying range of grass-roots publications that continues to expand. The rise of more sophisticated services will shape future development: if authors of electronic materials know that third party services can scan their documents for place names and automatically generate illustrative maps, for example, they will have a concrete incentive not only to write "Springfield, MO" (vs. Springfield with no determiner) but to make greater use of geographic detail.

Professional humanists have worked long and productively to establish common guidelines for encoding textual information. The Text Encoding Initiative's (TEI) guidelines for XML [14, 15] represent more than a decade of work and the TEI continues to explore new ways to represent semantic information. Architectures for standoff markup allow us to encode ever denser and more sophisticated data structures.[16-22] The Semantic Web is developing new mechanisms to exchange XML data.[23, 24]

Nevertheless, even the most sophisticated editors (often faculty funded by the NEH) and collection developers (often librarians funded by the IMLS) create vertically structured systems, in which style-sheets define the functionality and appearance of the print/PDF/HTML/RTF/etc. output that end users actually encounter. We have, as a community, only begun to consider the implications of complex digital library environments, and as language technologies continue to progress, entity detection and tracking, single and multi-document summarization, and cross-language information retrieval, and other services are likely to become standard components of our intellectual worlds. We need to understand the implications for our own work as we in the humanities create electronic sources on which teaching, learning and research will depend for many years to come.

Language technologies have both tactical and strategic implications. We see the rise of corpus editors who design documents to work with evolving systems and manage collections too large for conventional editing techniques (e.g., fifty years of a major newspaper)[25, 26]. At a tactical level, we need to imagine how our documents will interact with their electronic environments over time. There are various levels of detail in which we can, for example, apply TEI structured markup – in one five-level taxonomy, the fifth level is explicitly open-ended [27]. A digital library might contain hundreds of thousands of manually edited annotations, but automatic services such as morphological analysis, dictionary lookups, word clustering, citation aggregation, and keyword/phrase identification can add millions of additional links [28-34].

At a strategic level, emerging technologies will continue to challenge us to rethink the ways in which we develop collections [31, 33, 35-38]. The heavy use of automatic linking, for example, encourages the inclusion of keyword- and phrase-based encyclopedias, while city-scale GIS datasets make gazetteers of street names valuable. Evolving technologies such as cross-language information retrieval and statistically based machine translation enhance the value of bilingual lexica.

It is important, therefore, to design collections to work with services, both current and emerging, upon which users increasingly depend, and to engage in a conversation about how collection design and technological opportunities can better serve our various user communities. It will be important, as research proceeds in all these areas, to develop a broad and imaginative understanding of user communities, the kinds of information they want, and the information-seeking strategies they prefer to pursue. A detailed assessment would provide a roadmap for future language-technology research.

Technologies to be evaluated

Evaluation Metrics

The various technology evaluation forums have struggled to develop efficient mechanisms to compare the results of different systems when applied to the same collections [39, 40]. Many humanists, however, focus their research on topics that are not easily quantifiable: on the ambiguities inherent in human language, culture, and belief. It will be important to create a forum in which different communities can systematically explore how well existing evaluation metrics serve their research needs. It may be that semi-automated techniques that provoked suspicion among some users will be found, upon examination, to provide reasonable and usable results: in this case, publishing the results of this analysis may allow language technologists to convey their results to whole new communities. In other cases, new communities may spur changes in evaluation methodology. Earlier evaluations of summarization systems, for example, often compared machine-generated summaries with human-composed abstracts, but many humanists would object to the notion that source documents have unambiguous

messages that lend themselves to such techniques. As a result, the guidelines for the 2003 Document Understanding Conference [41] now attempt to address such problems by calling for summarizations that reflect particular points of view. The outcome will be to increase the sophistication and the understanding of evaluation metrics for various technologies.

In addition, there is also a need for establishing a new, digital library-specific evaluation track for Language Technologies (LT), in order to motivate LT researchers to start tuning their methods and tools to deal with the issues specific to cultural heritage collections. For example, in the context of entity detection and tracking, systems can be evaluated on how well they can use large-scale external gazetteers as those provided by existing digital libraries, e.g., the Alexandria Digital Library[3]. Another part of the problem is how the output produced by the LT systems will be disambiguated and merged with other existing DL resources, such as encyclopedia. For example, in order to link automatically "Cambridge" in the text to the correct encyclopedic entry, the EDT system needs to disambiguate this location (Cambridge, UK vs Cambridge, MA) – a task that was only recently addressed in the database track of ACE, which, however, is not likely to be continued due to lack of funding.

Visualizations

Good backend data analysis needs to be presented in a way that users can understand and build upon. Visualization strategies will constitute a theme underlying each individual technology and our attempts to synthesize disparate tools into coherent systems. Visualization strategies are part of many language technology and humanities computing projects. DARPA's *Command Post of the Future Project* [42] may have goals that differ widely from our user communities and assume a far more complex technical infrastructure that we can deliver to academics or amateurs, but programs such as this support major efforts to represent complex information extracted from textual and other sources about evolving events in time and space.

Source quality

It is important to document the correlation of data quality and retrieval vs. the costs of data entry for historical materials. Substantial research has evaluated information-retrieval strategies for errorful sources produced by automatic speech recognition and optical character recognition systems, showing, for example, that searching for n-grams (as opposed to words or stems) can improve retrieval for noisy data. We must expand this research to encompass challenges unique to the capture of historical materials. These include OCR of problematic texts including non-Roman alphabets such as classical Greek, early modern texts with problematic print, eighteenth century books that use the long s (e.g., "bleſsed" for "blessed," where the long-s resembles an f), multicolumn documents where page analysis engines run columns together. We should also examine the use of word-spotting techniques for handwriting recognition.

It will be important to assess when uncorrected (and thus inexpensive) OCR output provides acceptable performance [43]. Uncorrected OCR clearly suits the current needs of those working with large collections such as the *Making of America* digital libraries at Cornell and Michigan. Research on mining data from OCRd bilingual lexica [44] will probably provoke initial resistance from those working with cultural heritage languages.

In some cases, scholars are interested in words rather than in information: they are studying the semantics of particular terms (e.g., "shame" vs. "guilt"). Some in this group expect very high standards of precision and recall based on print concordances and, more recently, on carefully prepared text corpora. Discussion and analysis may, in fact, reveal that a larger corpus of uncorrected OCR supports much of their actual work better than smaller carefully edited corpora. Likewise, we will also be able to document and quantify current limitations on existing tools – limitations many users sense but are not in a position to document.

[3] State-of-the-art EDT systems avoid using large-scale gazetteers due to the very high levels of ambiguity that they introduce. However these resources are needed to ensure adequate coverage so recently some new research has started on these issues (see http://www.kornai.com/NAACL) but still substantial advances are needed.

Any evaluation study must bear in mind that language technologies and the corresponding expectations of users evolve together. Thus, a large collection of raw OCRd text may provide satisfactory results with current systems but the error rate may have cascading effects that degrade the performance of higher-level systems. We therefore need to consider the implications of source quality for every technology that we examine.

Web/XML Information Retrieval

Besides the quality of the source data, we must also analyze the costs and benefits of structured markup. International teams of humanists have, over the course of fifteen years, developed evolving guidelines for SGML and now XML markup, the Text Encoding Initiative (TEI [14]). Adding such markup increases the cost of text production. Many data producers, therefore, resist following the TEI guidelines and produce simple HTML. Furthermore, markup is open-ended: it ranges from a simple header followed by a stream of ASCII text to detailed linguistic encoding [27, 45].

Before we invest in manual text markup, we need to assess the services systems now, and in the foreseeable future, can provide with and without various levels of structured markup. Services that work with random web pages provide a baseline. Work on XML information retrieval – both database- and document-oriented – has expanded ([46, 47], [48]; almost fifty groups are participating in the 2002 Initiative for Evaluation of XML Retrieval [49]) and these efforts promise to let users exploit expensive structured markup.

In practice, we expect that smaller, carefully tagged collections will provide training sets that support the searching and analysis of much larger bodies of text. Determining the cost/benefit tradeoffs for varying levels of markup, however, requires analysis of another issue: to what extent can we automate various classes of markup? Most language technology tools add some sort of information to their collections – even simple vector-space IR systems calculate word frequencies, thus effectively adding significance tags to particular terms.

Entity/Relation Detection and Tracking:

In informal, though extensive, conversations, our colleagues stress their interest in being able to locate named entities such as people places, technical terms and things. They want to locate passages about Salamis in Cyprus (rather than the much more famous Salamis near Athens), instances where "York" designates the duke of York rather than the place – and, if possible, citations of a particular duke of York (rather than his father or son), references to particular dates (e.g., all information on events during the first Lincoln Douglas debate). Those who work with traditional historical tools understand and treasure effective indices. We have thus had little difficulty explaining the significance of entity detection and tracking.[4]

We need to explore the implications for differing usages to which the language technology community puts open-ended news-sources and humanists put their historical sources. Humanists devote immense amounts of effort to preparing and publishing source materials. Editors can spend years assembling information about major canonical literary works such as *Oedipus Rex* or *Hamlet* – even in print editions where space is a limiting factor, the scholarly apparatus can far exceed the size of the edition. Collections drawn from nineteenth century US

[4] Entity detection describes the ability to recognize when a named entity is first mentioned in a text: e.g., recognizing that the "Lee" in a passage is Fitzhugh rather than Robert E. Lee. Entity tracking describes the ability to link subsequent expressions such as "the general" or pronouns such as "he" and "his" correctly to Fitzhugh Lee or their proper referent [11]. Once these named entities have been identified, relationships between them can be discovered. Thus, if we have "Judge Worthington in the Middlesex District Court," we try to record not only references to "Judge Samuel P. Worthington, 1818-1877" and "Middlesex Court, Middlesex County, MA" but the fact that Worthington was a judge in this particular court. The Automatic Content Extraction competition is currently studying ways to identify and then automatically recognize a set of core relations: at present, the relations are Role, Located and Part [50]. In the example above, a Role relation would link judge to court.

newspapers, 18th century British court records, or even more specialized genres such as regimental histories from the Civil War are much larger than the most elaborate editions with which humanists are accustomed to work. We need to consider the challenges of supporting a community of scholars who spend years refining heuristics tuned for their particular corpora.

Consider the following scenario. A corpus editor decides to prepare an edition that adds value to digitized court records in five US cites from the 1850s. She assembles this collection and then submits a small subset of documents to a categorization system, which locates the most similar documents on which entity detection and tracking have been systematically applied. The editor then extracts the rules – which may be heuristics or statistical associations or both – and uses these to bootstrap her own work. She develops mechanisms to identify various forms of libel within her collection, but also uses these techniques to identify relevant texts from other collections as well (e.g., newspapers, diaries, etc.). Editors of newspaper collections then use her techniques for categorizing court cases to structure their work.

We need to evaluate the variety of entity-detection and tracking technology that is now emerging. Some systems, such as Sheffield's *GATE* [19], rely on human-generated rules, while others, like BBN's *Identifinder* [12, 13], apply learning algorithms to tagged texts. Cost/benefit tradeoffs may be very different for humanists, however. Where intelligence analysts, who need to adapt their systems to new topics and new streams of data, may prefer a system that was easily trained on new topics at the cost of somewhat lower accuracy (which generally favors an emphasis on training sets and stochastic methods), humanist editors will be much more likely to invest labor in developing heuristics in order to improve system accuracy.

Encoding guidelines also deserve close scrutiny, as they define the agendas which technologies pursue and incorporate assumptions about collections. The Entity/Relation Detection and Tracking guidelines for the Automatic Content Evaluation (ACE) group [50] serve the needs of that particular competition but they provide a natural starting point for other communities. Likewise, work on annotation temporal information [51, 52] represents a step forward in understanding what we can and cannot represent.

Document Understanding

We use this broad rubric more broadly than the Document Understanding Conference (DUC). The DUC 2003 evaluation [41] will focus on four carefully delineated tasks: generating very short summaries (c. 10 words), short summaries (= c. 100 words), short summaries representing a particular viewpoint, and short summaries in response to a question. We broaden this category to include any technologies that help users understand the content of aggregate documents and consider individual technologies subsumed within the above competition (e.g., single document summarization, question answering systems, clustering and categorization, etc.).

We see two complementary sets of questions. First, to what degree do the summarization problems as delineated in the DUC 2003 and subsequent evaluations serve the varied needs of our audiences? An individual exploring a city can be viewed as delivering a series of automatically generated, spatially based queries: "what do we know about my current location?" "to what extent does what we know about this location match my past interests?" These two queries come very close to tasks four and three in DUC 2003. Second, what types of summaries would users find useful? Users may well prefer a schematic summary to connected prose. If we have identified a reference to a particular person, the system might attempt to fill out a standard template (picture, birth and death dates, career summary) and then present what appear to be the ten most significant statements about this particular person. We must document the varying information needs of users and to provide the summarization community with data as to how they might develop their systems.

Other technologies that should be considered in conjunction with document understanding include:

Simple Single Document Summarization: Light-weight approaches to single document summarization examine subsets of a document, apply various heuristics to rank these document fragments and display the most important. Document subsets can be sentences, paragraphs or phrases.

Question answering systems: The DUC 2003 evaluations ask for summaries in response to a question. We will also consider question-answering systems that simply locate the answers to particular questions. The question-answering track at TREC has shown substantial progress,[53] with Q&A competitions becoming steadily more demanding. Summarization systems that track discrete facts such as birth and death dates of individuals are, in fact, question-answering engines that apply checklists of questions to collections as a whole. The Q&A systems thus deserve to be considered in their own right, both because of their own potential and because of their significance for higher level summarization processes.

Categorization and Clustering: Categorization systems attempt to match new documents to existing taxonomies. Thus, a categorization system might attempt to group documents according to the library of congress subject catalogue or, in more recent work, according to text genre.[54-57] Clustering algorithms look for patterns inherent within the data. Clustering is useful for those who wish to discover unknown patterns and associations.

While filtering and collaborative filtering are often treated separately, we would consider them as special cases of categorization. Filtering determine whether or not to alert users about a new document by comparing the documents against user profiles. User profiles can be manually created, inferred from past user actions or exploit both explicit and implicit data. Collaborative filtering systems cluster individual users with other groups of users, then use the aggregate profiles for the group to determine the relevance of new documents.

Likewise, one could, for the purposes of discussion, treat topic detection and tracking as a subset of the categorization/clustering problem. In topic detection and tracking, systems attempt to identify documents on a single theme (e.g., D. C. Sniper) and then determine whether new documents contain novel information or simply restate old news. While humanists are not filtering thousands of news sources during a national emergency in real time, topic detection and tracking does have clear applications for users tracking particular subjects (e.g., formal articles on staging in *Antony and Cleopatra* or Civil War enactor events in South Carolina).

Integrating Language Technology with the Digital Library: Need for Infrastructural Work

In order to allow DLs to benefit from Language Technology, these new components need to be integrated as part of the digital library itself, which has certain infrastructural implications. One aspect is making these new tools available to the DL users in an intuitive way and delivering their results over the Web. Another important aspect is *sharing* language technology tools, i.e., making them an integral part of the digital collection and allowing other digital libraries to use these tools on their collections. For example, the Perseus digital library already has extensive language resources for Greek, e.g., morphological analyzer, so the need is to enable another digital library that contains Greek documents to offer its users this service without having to install and run locally the tools as is the only possible approach at present. These issues are being addressed by research on architectures and infrastructures for language technology, e.g., the widely-used, open-source *GATE* system[19], and we would recommend a closer integration of such infrastructures into digital libraries.

Another important aspect of such infrastructural work is that it aims to make language technology easier to customize and use by non-computational users, e.g., provide Web-based tools for collaborative corpus annotation, improve the portability of systems between tasks and domains, allow non-specialist users to customize the linguistic data and algorithms. The user communities discussed here can play a vital role in evaluating the existing solutions and making their needs known to the technology developers and only through such future collaboration the necessary language technology tools for cultural heritage applications can be created.

Citations

1. 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.
2. Rosenzweig, R., *Everyone a Historian -- afterthoughts to the Presence of the Past*. 2002, George Mason University.
3. *United States Historical Society Directory*. 2002, D'Addezio.
4. Listokin, D. and M.L. Lahr, *Economic Impacts of Historic Preservation*. 1997, New Jersey Historic Trust: Trenton, NJ. p. 484.
5. Leithe, J. and P. Tigue, *Profiting from the Past: the Economic Impact of Historic Preservation in Georgia*. 1999, Athens, GA: Georgia Historic Preservation Division: http://www.gashpo.org. 26.
6. Listokin, D., et al., *Economic Impacts of Historic Preservation in Florida*. 2002, Florida Department of State, Division of Historic Resources, Bureau of Historic Preservation.
7. Commission, M.H., *2002 Massachusettts Historic Preservation Conference, Friday, September 27, 2002: Preservation Works: the Economics of Preservation*. 2002.
8. Burns, K., et al., *The Civil War*. 1989, PBS Video: Alexandria, VA.
9. Toplin, R.B., *Ken Burns's The Civil War : the historian's response*. 1996, New York: Oxford University Press. xxvii, 197.
10. 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.
11. 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.
12. Bikel, D., et al., *Nymble: a High-Performance Learning Name-finder*. Proceedings of the fifth ACM conference on Applied Natural Language Processing, 1997: p. 194-201.
13. Technologies, B. and AFRL/IFED, *Information extraction (IE) technology for counterdrug applications*. 2001, Department of Defense: Counterdrug Technology Development Program: Washington, DC. p. 5.
14. Sperberg-McQueen, C.M. and L. Burnard, eds. *TEI P4: Guidelines for Electronic Text Encoding and Interchange -- XML-compatible version*. 2001, TEI-Consortium.
15. Sperberg-McQueen, C.M. and L. Burnard, *Guidelines for the Encoding and Interchange of Machine-Readable Texts*. Version 1.0 ed. 1990: The Associate for Computers and the Humanities; the Association for Computational Linguistics; the Association for Literacy and Linguistic Computing.
16. 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.
17. Grishman, R. and T.P.I. Contractors, *TIPSTER Text Architecture Design*. 1998, New York University: New York. p. 70.
18. Program, A.C.f.t.T.T.P.I., *TIPSTER Text Phase II Architecture Concept*. 1996.
19. Cunningham, H., et al., *Developing Language Processing Components with GATE (a User Guide)*. 2002, The University of Sheffield: Sheffield, UK.
20. 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.
21. Bird, S. and M. Liberman, *A formal framework for linguistic annotation*. Speech Communication, 2001. **33**(1,2): p. 23-60.
22. 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.
23. Miller, E., et al., *W3C Semantic Web*. 2001, W3C World Wide Web Consortium.
24. Berners-Lee, T. and E. Miller, *The Semantic Web lifts off*, in *ERCIM News: online edition*. 2002.
25. 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.

26. 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.
27. Friedland, L., et al., *TEI Text Encoding in Libraries: Draft Guidelines for Best Encoding Practices (Version 1.0)*. 1999.
28. Crane, G., *New Technologies for Reading: the Lexicon and the Digital Library.* Classical World, 1998. **92**: p. 471-501.
29. Crane, G., *The Perseus Project: An Evolving Digital Library*. 2000, Tufts University.
30. Crane, G., *Designing Documents to Enhance the Performance of Digital Libraries: Time, Space, People and a Digital Library on London.* D-Lib Magazine, 2000. **6**(7/8).
31. Crane, G., et al., *The symbiosis between content and technology in the Perseus Digital Library.* Cultivate Interactive, 2000. **1**(2).
32. Crane, G., et al., *Drudgery and Deep Thought: Designing Digital Libraries for the Humanities.* Communications of the ACM, 2001. **44**(5).
33. 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.
34. 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.
35. Crane, G., *In a digital world, no books 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.
36. Crane, G. *Cultural Heritage Digital Libraries: Needs and Components*. in *European Conference on Digital Libraries*. 2002. Rome: Springer.
37. Crane, G., *The Perseus Project and Beyond: How Building a Digital Library Challenges the Humanities and Technology.* D-Lib Magazine, 1998.
38. Crane, G., *Extending a Digital Library: Beginning a Roman Perseus.* New England Classical Journal, 2000. **27**(3): p. 140-160.
39. Voorhees, E.M. *Overview of TREC 2001*. in *TREC 2001*. 2001. Gaithersburg, MD 20899: NIST.
40. *ACE Evaluation plan version 06*. 2002.
41. Buckland, L., *DUC 2003: Documents, Tasks, and Measures*. 2002, National Institute for Standards and Technology.
42. Page, w., *Command Post of the Future*. 2002, DARPA.
43. Darwish, K. and D.W. Oard. *Term Selection for Searching Printed Arabic*. in *SIGIR 2002: The Twenty-Fifth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2002. Tamere, Finland: ACM Press.
44. Mayfield, J. and P. McNamee. *Converting on-line Bilingual Dictionaries from Human-Readable to Machine-Readable Form*. in *SIGIR 2002: The Twenty-Fifth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2002. Tamere, Finland: ACM Press.
45. Hockey, S.M., *Electronic Texts in the Humanities: Principles and Practice*. 2001, Oxford and New York: Oxford University Press.
46. Fuhr, N., N. Gövert, and K. Großjohann. *HyREX: Hyper-media Retrieval Engine for XML*. in *SIGIR 2002*. 2002. Tampere, Finland: ACM.
47. Abolhassani, M., et al., *HyREX: Hypermedia Retrieval Engine for XML*. 2002, University of Dortmund: Dortmund.
48. 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.
49. Fuhr, N., M. Lalmas, and G. Kazai, *INEX: Initiative for the Evaluation of XML retrieval*. 2002, University of Dortmund.

50. *Automatic Content Extraction: ACE -- Phase 2 -- Documentation*. 2002, National Institute for Standards and Technology.
51. Ferro, L., et al., *TIDES Temporal Annotation Guidelines*. 2001, Mitre.org: McLean, VA. p. 57.
52. Pustejovsky, J., et al., *TimeML Annotation Guidelines*. 2002, Brandeis University: Waltham, MA. p. 49.
53. Voorhees, E.M. *Overview of the TREC 2001 Question Answering Track*. in *TREC 2001*. 2001. Gaithersburg, MD 20899: NIST.
54. Lee, Y.-B. and S.H. Myaeng. *Text Genre Classification with Genre-Revealing and Subject-Revealing Features*. in *SIGIR 2002: The Twenty-Fifth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2002. Tampere, Finland: ACM Press.
55. Stamatatos, E., N. Fakotakis, and G. Kokkinakis. *Text Genre Detection Using Common Word Frequencies*. in *COLING2000: the 18th International Conference on Computational Linguistics*. 2000. Saarbrücken.
56. Kessler, B., G. Nunberg, and H. Schütze. *Automatic Detection of Text Genre*. in *ACL 97: Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. 1997.
57. Rauber, A. and A. Müller-Kögler. *Integrating Automatic Genre Analysis into Digital Libraries*. in *JCDL 2001: First ACM/IEEE Joint Conference on Digital Libraries*. 2001. Roanoke, VA: ACM Press.