The Burlington Agenda: Research Issues in Intellectual Access to Electronically Published Historical Documents

Elizabeth H. Dow, with David R. Chesnutt, William E. Underwood, Helen R. Tibbo, Mary-Jo Kline, and Charlene N. Bickford

Abstract

As increasing numbers of primary historical documents appear on the World Wide Web, publishers of those documents will need ways to provide intellectual access to the contents. At a three-day meeting in Burlington, Vermont experts in experimental electronic publishing, library and information science research, documentary editing, and computer science research identified (1) the need for user studies to determine the needs and reactions of the audience(s), (2) the need to assess implications for change in publication management, and (3) the need to compare empirically various technological approaches to access to information as three areas of research that can contribute to our understanding of how to construct and improve intellectual access to historical documents on the Web.

Between December 1997 and December 2000, the Special Collections department of the Bailey/Howe Library at the University of Vermont (UVM) developed the "George Perkins Marsh On-line Research Center."[1] Through the Center the university provides, on line, more that 600 primary documents by, to, or about nineteenth-century linguist, naturalist, and diplomat George Perkins Marsh, chosen from archival collections at UVM, Harvard, and the Smithsonian Institution. The department saw this project as a way to share with scholars throughout the world, popular portions of one of its most heavily-used collections. It argued to the funder, the Woodstock Foundation (a small, private foundation in Marsh's hometown of Woodstock, Vermont), that the project would provide researchers access to

[1] *Burlington Agenda,* 2000. <http://etext.uvm.edu>, September 29, 2001.

the documents through their desktop computers, enabling a number of scholars to simultaneously study (and perhaps discuss) a single document, regardless of the scholars' location. Furthermore, the documents themselves receive less handling when viewed on-line than in the reading room.

The original grant paid for diplomatic transcription of the documents, which are primarily letters. Recognizing that many people who could use the letters might not understand many of the references contained in them or the context of their creation, Special Collections proposed, and the Woodstock Foundation funded, a project extension that enabled the department to add footnotes and essays to clarify the documents' contents and contexts. The transcription is linked to a scanned image of the original, if it is owned by UVM; all transcriptions link to the EAD, published inventory for the full Marsh collection. At the completion of the project, the collection had all the editorial apparatus of a modern documentary edition except for an "index." For in-depth intellectual access, it had only the full-text search engine built into its publication software (DynaText™ and DynaWeb™). The need to develop an index equivalent for this electronic scholarly edition drove the efforts that culminated in the Burlington meeting, which produced the research agenda summarized in this article.

The Meeting

From April 7 through 9, 2000 ten experts[2] from the fields of experimental electronic publishing, library and information science research, documentary editing, and computer science research, convened in Burlington, Vermont to discuss ways to improve and standardize intellectual access to electronically published historical documents. The meeting, sponsored by the University of Vermont and funded by a grant from the National Historical Publications and Records Commission, focused on identifying research issues that will contribute to an intellectual framework or set of editorial guidelines for publishing historical documents in a way that assures effective intellectual access.

The Need

The World Wide Web provides a highly attractive distribution medium for primary historical documents, and hundreds of repositories around the world have begun publishing historical documents there. To enable scholars and nonscholars alike to retrieve those documents effectively, web publishers need to know how to provide intellectual access to their documents, i.e., to assure that potential users can find documents through what may involve several layers

[2] See the appendix for the names of participants.

of discovery. A user must first find the web site that holds relevant materials. Then the user must find the right document, or even a phrase or passage within a document. It is not a trivial problem, and the current growth rate of the Web guarantees that it will not diminish.

Modern printed editions of historical documents provide a highly sophisticated level of information retrieval through back-of-the-book indexes. The indexers of these volumes, usually the editors, capture references ranging from ordinary terms such as "rivers" to sophisticated abstractions such as "human rights." They sort out people and places that have the same names. They organize and classify the contents of the documents. They aggregate concepts presented in many different ways by analyzing the meaning of the text and labeling the concepts with a consistent vocabulary. They frequently use cross-references. Retrieval is precise and complete, but the editorial work is highly labor-intensive. Publishers of historical documents on the web need to provide access to their document contents with an efficiency and accuracy that will rival or exceed the standard set by the indexes of printed historical documentary editions, in a more cost-effective way.

Provision of high-quality intellectual access to electronically published historical documents is difficult because of (1) the nature of primary documents, (2) the traditions of modern indexing, and (3) the very "webness" of the World Wide Web. Primary historical documents are full of oblique and unidentified references, undefined terms, abbreviations, and words that have many meanings depending on discipline, time, and context. A concept may be discussed throughout a lengthy document with no specific reference to that concept in language recognized by modern readers. All of these linguistic and conceptual complications may be published without the supporting clarification or analysis customarily found in secondary literature. In an edited collection of documents, the identification of people, places, and topics may appear in the annotations, but many, if not most, abstractions appear only in the index. Unedited documents lack even rudimentary clarification of the simplest factual matter.

Traditional book indexing works within a closed system, i.e., the volume or set of volumes it references. Each index is physically bound with its book; the two will never exist independent of each other, nor will the index ever be consulted in reference to anything else. The look and feel of indexes have been standardized to a narrow range of structures that most scholars understand, having learned to use them as school children. Like all book indexers, documentary editors practice the time-honored tradition of literary license and develop their own systems including their own thesaurus of index terms, their own standards of detail, and their own method for handling abstraction, ambiguities of language, etc., for a given set of volumes. They tend to see their documents as both the source of problems and the source of many of the solutions they develop. However, the time needed to edit and index voluminous sets of documents works against consistent indexing. Even in very stable scholarly

environments, staff changes disrupt indexing consistency.[3] Even when one person does all the indexing for a whole collection, both the thesaurus of terms and the use of terminology evolve as the interests of researchers change over time. Editors rarely have the luxury of time and resources to go back and rework their indexes.

Publishers of historical documents have become less and less able to predict who will use their materials because the Web opens historical documents to a wide range of ages, educational levels, and ethnic and national groups. Given the diversity of the audience, no editorial apparatus will serve all Web audiences. Worse, if a publisher deliberately targets a specific audience through the Web, s/he has no generally accepted way of presenting editorial apparatus—no generally accepted "system" like a back-of-the-book index—the audience knows and can use. It is very unclear what the "index" of a collection of electronically-published historical documents, with or without editing, might look like.

Further, web search engines can flatten or destroy the editorial structure a publisher develops within a collection of documents. Formal documentary editing contextualizes documents, which is to say, it orients or situates documents among other historically and intellectually related documents. Frequently (and by design) the order in which documents and added apparatus appear in the book contribute to that contextualization. In the Web environment, however, a publisher can neither predict nor control the level at which any one user will "discover" material. Web search engines find materials at the level comparable to the "front" of a collection as well as at the level of an individual document.

Today's search engines can find explicit terms that occur in a text; they can find terms that the user links together; they can isolate terms from other terms. Inverted indexes, field-specific indexes, and search algorithms based on statistical analysis of the language of the documents can retrieve items that reflect the concepts in the search. Indeed, a superficial examination of a printed index would reveal that search engines available at this writing could retrieve many of its references from the full text of a work, but it is also clear that they could not build the web of relationships and analysis provided by a good back-of-the-book index, nor the navigational support provided by such an index. While search engine developers consistently have improved their products over the past forty years, we still do not know what technology can eventually do, what will always requires human intervention, and how the two can be woven into useful models for providing intellectual access to electronically published historical documents. It is an area begging for research.

[3] The Papers of Thomas Jefferson have produced 32 volumes of edited documents over the past 51 years. The editors estimate that they have at least 22 volumes yet to publish. Caroline Moseley, "27 volumes—and 20 more to come." *Princeton Weekly Bulletin,* May 10, 1999, <http://www.princeton.edu/pr/pwb/99/0510/jeffsn.htm>, September 29, 2001.

The Environment

Intellectual access to web-published historical documents occurs in an environment which includes the technology of electronic publishing, the emerging world of metadata, and a wide variety of information retrieval approaches. The following paragraphs provide a brief general explanation of the environment in order to provide the reader with enough background to understand the research issues presented later. They do not constitute either a comprehensive literature review or comprehensive coverage of the problems under discussion.

Electronic Publishing

In an effort to provide a data format that will remain useable regardless of the hardware or software employed, electronic publishers on the Web have turned to the Standard Generalized Markup Language (SGML), or a modified version of the SGML standard called the Extensible Markup Language (XML). SGML defines the rules for creating text markup through sets of tags, but does not dictate which tags the sets should contain. Different types of document types require different tag sets, so a tag set defined for a particular type of document is referred to as a "document type definition" or DTD. To date, most web publication has used the Hypertext Markup Language (HTML), a DTD for web pages. HTML does not adequately represent the complexity of more highly structured documents, which prompted the development of Encoded Archival Description (EAD) as a DTD for archival inventories; the Model Editions Partnership (MEP) has developed a DTD for historical documents and their editorial apparatus.[4]

Regardless of the DTD, markup serves several purposes. It defines the structure of the document (e.g., head, salutation, body, closing, signature); it clarifies the nature or source of the text (e.g., person, ship, supporter, supplied); and it links one document to another or to information outside the document itself (e.g., ixe (index), xptr, xref).[5] Each tag may also contain one or more modifiers, called an *attribute*. Using attributes, the publisher can particularize the content of a tag. For example, in May 1862 George Perkins Marsh confided

[4] A standard SGML or XML tag contains an element name in angle brackets, for example, <person>, which precedes the text it refers to. At the end of the text, the tag is closed by a set of angle brackets with a slash before the element name, e.g., </person>. For a theoretical explanation of the use of markup languages, see J. H. Coombs, Allen H. Renear and Steven J. DeRose, "Markup systems and the future of scholarly text processing." *Communications of the ACM*, 30 (November 1987): 933–47. For a discussion of how markup can enhance electronic publication of historical documents, see David Chesnutt, "Model Editions Partnership: Smart Text and Beyond." *D-Lib Magazine* (July 1997) <http://www.dlib.org/dlib/july97/07chesnutt.html>, September 29, 2001.

[5] All tag examples come from the MEP DTD. Model Editions Partnership Reference Manual <http://adh.sc.edu/meptsdv1.html>, September 29, 2001.

to his friend Hiram Powers[6] that "I am mighty nervous about Yorktown. I am not one of those who put their trust in the 'quaker general,'" referring to General George B. McClellan's conduct of the Peninsula Campaign during the American Civil War. In marking up that letter, the publisher might choose to modify Marsh's nickname for McClellan through a "regularizing" attribute and produce the following tag: <person reg="McClellan, George Brinton">quaker general</person>. With comprehensive use of tagging and attributes, markup can support intellectual access by identifying the function of a fragment of text, by normalizing names, making words less ambiguous, embedding index terms, and linking to external resources. Search strategies which specify a term within a specific tag or attribute significantly narrow and focus the results of the search.[7]

In addition to carefully considered decisions about markup, web publishing also involves basic elements of design. As the design of a book can effect its ease of use and its clarity, so can the design of a web site.[8] Whether and how the site presents an on-line equivalent to a table of contents, how the site is designed to plan and direct the user's navigation among its features, the degree to which the site includes flashy elements generally referred to as "eye candy"—these elements all contribute to a site's usability and therefore become part of the problem of intellectual access.

Metadata

Cultural repositories of all types have developed formal and informal traditions and structures for describing their holdings. Today, these descriptions have come to be called *metadata,* usually defined as "data about data." Broadly speaking, metadata involves any information about a document or file that describes its structure or content. Catalog cards, MARC records, archival finding aids, and museum registers and inventories can all be labeled "metadata," as can SGML markup.[9]

Most SGML files carry metadata about the file itself, and the document it includes, which usually appears in a large section at the top of the file, called

[6] George Perkins Marsh to Hiram Powers, May 1, 1862 <http://bailey.uvm.edu:6336/dynaweb/woodstock/gpmhp620501/@GenericBookView>, September 28, 2001.

[7] To distinguish one Queen Elizabeth from another, an editor would tag one <person reg="Elizabeth II, Queen of Great Britain, 1926–">Queen Elizabeth</person> and the other <ship>*Queen Elizabeth*</ship>.

[8] A great amount of guidance in website design appears on the Web. A good collection of sites has been developed by Yahoo!, Inc. <http://dir.yahoo.com/Arts/Design_Arts/Graphic_Design/Web_Page_Design_and_Layout/>, September 30, 2001.

[9] Jessica Milstead and Susan Feldman, "Metadata: Cataloging by any other name . . ." *On-line* 23 (January/February 1999): 24-6+. Also available on-line at <http://www.on-lineinc.com/on-linemag/OL1999/milstead1.html>, September 29, 2001. The research and development around metadata schema are in great flux. The collection of links to projects and organizations maintained by the International Federation of Library Associations stays as current as any. <http://www.ifla.org/II/metadata.htm>, September 29, 2001.

the *header*. The header doesn't appear on users' screens, but a search engine can find and react to it. Thanks to digital/network technology, cross-institutional access to historical evidence held in all types of cultural heritage repositories flutters on the horizon. Such access is highly desirable, because it will support more efficient research. Additionally, such access opens historical documents and exhibits to potential new users, in particular, students at all levels of education. However, creating links and effective intellectual access among them will not happen easily.

The advent of the Web has stimulated much interest in methods for storing descriptive data in document headers for web search engines to find and use. Several research projects are testing ways to create interoperability among the metadata schemes developed by different professions and institutions.[10] While libraries have a long history of standardized metadata schemes of many varieties, and archivists have begun to standardize many of their descriptive practices, other holders and publishers of primary documents have not.[11]

Long before computers, documentary editorial projects developed procedures for indexing that included project-specific thesauri, rules for treatment of problems encountered routinely from volume to volume, patterns of cross-reference, etc. Editors had no reason to develop firm standards for data structure, data content, or data value for back-of-the-book indexes. Instead they relied on the tradition of established "best practice" approaches.[12] Unfortunately, lack of true standards will prove problematic in an electronic union collection, holding many independently created collections.

Similarly, museums of all types have developed "best practices" in their descriptive work. They have also developed "best practice" approaches for creating web exhibits which vary by both repository focus and the implementation of individual institutions, and, they take justifiable pride in the style and creativity each has displayed. While each repository's web exhibitions or published collection will continue to have idiosyncratic needs, conceptual or actual union

[10] In addition, the efforts to create automated links between and among schema—popularly called "crosswalks"—can be tracked through a directory created by Google <http://directory.google.com/Top/Reference/Libraries/Library_and_Information_Science/Technical_Services/Cataloguing/Metadata/Crosswalks/>, September 29, 2001.

[11] Among examples of descriptive standards are: *Anglo-American Cataloging Rules,* edited by Michael Gorman and Paul W. Winkler, 2nd ed., 1988 revision. Chicago: American Library Association, 1988; MARC Standards Office, Library of Congress, *MARC 21 Format for Bibliographic Data,* 1999 ed. (Washington, D.C.: Library of Congress, Cataloging Distribution Service, 1999); *Library of Congress Subject Headings.* (Washington, D.C.: Library of Congress, 1975); Stephen L. Hensen, *Archives, Personal Papers, and Manuscripts: A Cataloging Manual for Archival Repositories, Historical Societies, and Manuscript Libraries* (2nd ed.), (Chicago: Society of American Archivists, 1989); *Encoded Archival Description* Application Guidelines, v. 1.0. (Chicago: Society of American Archivists, 1999) and *Rules for Archival Description,* (Ottawa: Bureau of Canadian Archivists, 1990).

[12] Mary-Jo Kline, *A Guide to Documentary Editing,* 2nd ed. (Baltimore: Johns Hopkins University Press, 1998); Richard N. Sheldon, *Readings in Documentary Editing: a Highly Selective List Compiled for the NHPRC Institute for the Editing of Historical Documents.* (Washington, D.C.: National Historical Publications and Records Commission, 1995).

databases of publications, primary documents, documentary editions, and art and artifact exhibits will require some standardization and uniform practice, although it is not well understood how much.

Information Retrieval

Most information retrieval systems, from card catalogs to on-line search engines, are *document* retrieval systems, i.e., they attempt to identify documents that address a user's request. *Passage* retrieval systems attempt to identify relevant passages within a document and indicate the sentence(s) or paragraphs related to a user request. Passage retrieval systems come closer than document retrieval systems to providing intellectual access that compares to a back-of-the-book index.[13]

To date, most information retrieval systems have depended on manipulating the words of the documents while ignoring any "meaning" they carry. Some document and passage retrieval systems integrate knowledge of the morphology of words,[14] the syntactic structure of sentences, and the semantic relationships of terms. They create a semantic representation of the user's query which they match against a similar representation of the text. These approaches have had mixed results.[15] None seems to have achieved the precision of a back-of-the-book index, but no research exists that addresses that question.

Researchers in natural language processing of text have observed that what makes text retrieval so difficult is the large volume of real-world knowledge not included in the text, but which is necessary to understand it. In an attempt to provide the needed background knowledge researchers have developed representations of the knowledge, in the form of conceptual specifications called *ontologies*.[16] In addition to the "broader," "narrower," and "used for" concepts built into common thesauri, sophisticated ontologies include many detailed characteristics related to a term. For example, the concept "ship" would include types of ships, the fact that ships have captains, the fact that ships may carry pas-

[13] G. Salton, J. Allan, and C. Buckle, "Approaches to passage retrieval in full text information systems." in *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, June 2–July 1, 1993, Pittsburgh*: 49–58; M. Kasakiel and J. Zobel, "Passage retrieval revisited." *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval July 27–31, 1997, Philadelphia,* 178–85.

[14] Word meaning; change caused by a change of tense, addition of a prefix or suffix, etc.

[15] D. D. Lewis and Karen Spark Jones, "Natural language processing for information retrieval." *Communications of the ACM,* 39 (1996): 92–101; T. Strzalkowski. F. Lin and J. Perez-Carballo, "Natural language information retrieval." *The Sixth Text REtrieval Conference (TREC 6)*. (Washington D.C.: NIST Special Publication 500-240, 1997): 347–66 <http://trec.nist.gov/pubs/trec6/t6_proceedings.html>, September 30, 2001; J. Ambroziak and William A. Woods, "Natural language technology in precision content retrieval," *Proceedings of the International Conference on Natural Language Processing and Industrial Applications,* (NLP+IA 98), August 18–21, 1998, Moncton, New Brunswick, Canada. <http://www.sun.com/research/techrep/1998/abstract-69.html>, September 29, 2001.

[16] T. R. Gruber, "What is an ontology?" <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>, September 29, 2001.

sengers, cargo, or weapons, the names of ships, the parts of ships, etc. If a passage retrieval search engine had access to this sort of knowledge to compare with the a body of text being searched, it could identify phrases like "the *Titanic*'s captain, Edward John Smith . . ." as relevant to a query about ships' captains, although the *Titanic*, is not identified in the text as a ship. The application of lexicons (machine-readable dictionaries) and thesauri—each a kind of ontology—to the problem may be a promising area to investigate. Obviously, the development of ontologies of terms and concepts specific to the historical context must precede natural language analysis of electronic historical editions. For instance, an ontology for a collection of papers from the period of the American Revolution would need to include knowledge of British provincial and American revolutionary government, social expectations and customs, economic and military apparatus and procedures, and other subjects.[17]

The Agenda

The *Burlington Agenda*[18] focuses on the need to address intellectual access to electronically published historical documents for the purpose of developing a widely applicable set of policies, practices, methods, and computer applications for use in publication environments ranging from those developed by highly sophisticated documentary editing projects to projects originating in small archives and special collections. Therefore, it emphasizes that the research

- should anticipate technological trends, but should not wait for them to materialize;
- should be interdisciplinary enough to draw from and build on knowledge available in a number of fields confronting similar issues, including artificial intelligence, information retrieval, human computer interaction, information extraction, and knowledge management; and
- should include a recognition that resources in technological expertise, funding, and even cultural/political power are limited within the community of publishers of historical documents. Thus researchers must find ways to coordinate activities to maximize available resources.

[17] The theories and practices in information retrieval have generated an extensive literature. Charles T. Meadow, Bert R. Boyce, and Donald H. Kraft. *Text Information Retrieval Systems.* 2nd ed. San Diego: Academic Press, 1999, provides an organized overview. Karen Sparck Jones and Peter Willett, eds., *Readings in Information Retrieval.* San Francisco: Morgan Kaufmann, 1997, pulls together classic articles in the field. The final article in that collection, Don R. Swanson's "Historical Note: Information retrieval and the future of an illusion" first appeared in the *JASIS* (39 (1988): 92–98) raising the issue of the limitations of automated retrieval. Robert Fugmann's piece, "Obstacles to progress in mechanized subject access and the necessity of a paradigm change," in William J. Wheeler, ed., *Saving the Time of the Library User through Subject Access Innovation: Papers in Honor of Pauline Atherton Cochrane.* (Champaign, Ill.: University of Illinois, 2000), addresses the issue at length and in detail. Chris Welty and Nancy Ide "Using the right tools: enhancing retrieval from marked-up documents." *Journal of Computers and the Humanities.* 33 (April 1999): 59–84, demonstrates the use of ontologies with historical documents.

[18] *The Burlington Agenda.* <http://etext.uvm.edu>, September 30, 2001.

The participants at the Burlington meeting recognized the spectacular potential for the web-based networking of cultural heritage resources that would deliver, "on demand": edited and unedited archival documents; sound archives of music, oral histories, and historical events; images, both moving and still, of people, places, events, arts, crafts, and common artifacts. One research issue addresses the network potential directly, but the language of the rest of the agenda reflects the meeting's roots in the world of the inscribed document. The authors found that repeatedly referencing all potential users created awkward writing and laborious reading, and so they decided to write the document from the perspective of the need that initiated the meeting. The language should not be construed as intending to exclude the needs of nondocument-based cultural heritage repositories; most issues apply to all.

Providing intellectual access to electronically published historical documents requires research in three very broad areas; user studies, publication management studies, and studies of access to information.

User studies: The Web makes a growing number of edited and unedited historical documents accessible to many audiences, ranging from the traditional readership of sophisticated scholars to school children all over the world. User studies should explore the information needs and behaviors of a wide range of users so publishers may optimize surrogation, presentation, user interface design, and other tools to facilitate intellectual access. Publishers cannot know how to provide intellectual access most effectively until they know how users hunt for materials, formulate questions, and evaluate search results.

Publication management: Publication management studies should explore procedural and management best practices in the creation and publication of both edited and unedited collections of historical materials.

Access to information: Using electronically published historical documents depends on procedures which the meeting participants labeled as "discovery, navigation, and retrieval." Discovery refers to the process by which potential users discover materials, i.e., finding the web site(s) that will supply useful information. Navigation refers to the process by which users move around in or among the web site(s). Retrieval refers to the process of identifying and collecting detailed information within a given web site. Navigation implies the use of links to move from page to page or document to document, whereas retrieval implies the use of a search engine within the web site to locate specific documents or passages within documents. Access to information studies should explore ways to improve these processes, understanding that, while it is convenient to think of discovery as a search of descriptors at the collection or exhibit level and retrieval as a detailed search within a collection or exhibit, in actual practice, the distinction becomes blurred. It is important to begin thinking about ways to improve the discovery of related collections, regardless of repository type and metadata traditions.

The participants in the Burlington meeting put forward the following eight issues as significant to the development of efficient and effective methods for providing intellectual access to historical documents. The issues contain within their broadly stated questions many specific research problems for individuals and institutions to address through projects involving a wide variety of disciplines and approaches. The *Burlington Agenda* includes a number of specific research projects and strategies that will address each issue, but does not begin to include all the possibilities. Large collaborative institutional efforts and funding arrangements can significantly enhance the research effort. However, well-conceived and developed small projects can also contribute answers to specific questions. Finally, none of these issues is completely new. Work exists on all of them in other domains. The *Burlington Agenda* serves to bring them into the domain of historical repositories.

Research Issue 1

Who are the users of electronically published historical documents, what do they need from sites that publish documents, and what do they need from the documents themselves?

Print editions of historical documents focus on the needs of the scholarly community and are generally available only in research libraries. The Web brings this work to everyone. Most existing electronic editions have been designed on the print metaphor, and without study of users and potential users. Through user studies, publishers can identify their primary, and perhaps secondary, audience(s). While no publication project can address the needs of every potential audience, many may find that, with relatively little additional effort, they can improve their publications to meet the needs of many more users than they have addressed in the past.

Research Issue 2

How do we assure that users effectively and efficiently navigate and retrieve information from collections within sites? In other words: how do we accommodate user needs through contextual and navigational aids?

While on the Web, users may come to a collection of documents along a path that does not recognize or use the editorial structure. They may not know that the search engine has brought them to the "middle" of a larger site, and that they would benefit greatly from various forms of editorial apparatus that the site may provide. Maintaining context in the electronic environment may require special apparatus or techniques. Web publishers need to know how users make use of navigational and other contextual tools before they can design effective sites.

Research Issue 3

What represents the most efficient and effective use of markup of electronically published historical documents to support intellectual access to text, to generate an effective presentation, and to provide contextual understanding during the discovery, navigation, and retrieval processes?

Whether a stand-alone item or part of a large collection or union database of documents, each web-published document will need some level of markup. Publishers need guidelines for the kinds and amount of markup they should use to create maximum intellectual access and effective presentation of information at minimum cost.

Research Issue 4

What constitutes the optimal range of uniform encoding practice and "indexing"[19] required to support effective discovery and navigation of items in a union database or collaborative efforts involving several types of cultural heritage institutions?

The promise of union databases of primary documents, cross-database searching, and multiple-site searching has fueled much of the effort to publish on the Web. Creating systems that assure that one publisher's data will play nicely with other publishers' data stands out as a major task. This issue can be addressed, in part, by developing: (1) guidelines for the effective markup and vocabulary enhancement for a union database, (2) guidelines for effectively establishing links between diverse cultural heritage collections through shared metadata, and (3) effective technologies for linking related materials across many institutions and institution types.

Research Issue 5

What are the benefits and challenges of linking electronically published historical documents to external resources, thus maximizing the use of available resources published on the Web?

The added editorial and scholarly information in well-edited historical documents requires a substantial amount of time and personnel to research, verify, annotate, and edit. Much of that time is spent in identifying people, places, and events for which shareable web-based information may already exist. The editing process could be made both more efficient and more accurate if documents or portions of documents could simply link to these resources and either display them as an enhancement to the document itself, or download the information into the document.

[19] Throughout the meeting, the participants used the term "index" and "indexing" to mean a wide variety of approaches to providing intellectual access.

Research Issue 6

What are the capabilities and limitations of information retrieval technologies for providing intellectual access to on-line historical editions? Can human intervention during the editorial process be used to overcome some of the limitations?

Traditionally, book-based editions of historical documents provide a level of intellectual access no search engine can rival. Book editors have at their disposal a table of contents, various lists, and, most powerfully, a back-of-the-book index, which is a kind of conceptual taxonomy for the book. In other words, the back-of-the-book index can be viewed as an ontology for the subjects of a document. Techniques in natural language understanding and information retrieval may ease the task of creating intellectual access tools by drawing on the existing ontological resources in published documentary editions. Even so, there remains a need to

- find a delineation of the point at which human effort (e.g., traditional indexing) is indispensable for optimal intellectual access within the current technological environment;
- discover the most effective automated processes for intellectual access; and
- create an experimental "baseline" vocabulary drawn from the documents themselves that can be applied retrospectively to other collections to improve computer-assisted subject analysis and enhance subject-based retrieval of documents.

Research Issue 7

How do we assure that users can find and retrieve information from within sites most efficiently and effectively, while retaining the contextual integrity of a set of hits within either a single database or union database?

A back-of-the-book index shows the researcher the language applied to define the volume(s) in use. Indeed, library research shows that people recognize the terms they need more quickly and accurately than they can predict them.[20] Given that electronically published historical documents may receive some terminological augmentation, and given that union databases of such collections will contain many thousands of documents, users may need to choose terms from some form of site-specific "index" before doing a search, and may need something other than a list of hits to work with after they have done the search.

[20] Marcia J. Bates, "Indexing and access for digital libraries and the Internet: human, database, and domain factors." *JASIS* 49 (November 1998): 1185–205; Susan Siegfried, Marcia J. Bates, Deborah N. Wilde: A Profile of End-User Searching Behavior by Humanities Scholars: The Getty On-line Searching Project Report No. 2. *JASIS* 44 (June 1993): 273–91; Marcia J. Bates, "Rethinking Subject Cataloging in the On-line Environment." *Library Resources and Technology Services* 33 (October 1989): 400–12.

Research Issue 8

What kind of publication practices will be appropriate for editors of electronically published historical documents in order to assure that they meet the users' need to find information, as well as meeting their goal of providing suitable context for that information.

Indexing historical documents published in print has always been, and remains today, an idiosyncratic endeavor. Incorporating knowledge-based tools into the editing process has the potential to drastically change this endeavor. Editors, and those who fund documentary publishing projects, need to know how the process will change. They need guidelines for cost-effective procedures for publication projects, a greater understanding of staffing patterns and the knowledge required for the electronic publication environment, and new process models for documentary editing and/or publishing.

Resources

Resources come in the form of individuals and institutions with whom researchers can collaborate and from whom they may receive funding. Collaborators can also provide contacts with funders, political support, test beds of documents, test populations, counsel and advice, expertise in research design and evaluation, trained staff time, and proprietary technology. Such collaborators might include:

Repositories of all sorts that collect, organize, and disseminate primary documents and other forms of evidence relating to our cultural heritage, e.g., special collections departments in college and university libraries, presidential libraries, governmental archives, and a wide variety of museums and galleries. All these types of repositories are publishing their holdings on the Web; all want their work discovered and used. All have much to learn from each other.

Educational institutions and associations, such as state departments of education and individual schools and teachers using the web as an educational tool. These institutions want well-designed Web sites to meet curriculum requirements. Furthermore, these institutions understand how they and their students use the Web and the resources they find there today, and can provide information on how use of the Web changes over time.

Academic researchers in conceptual taxonomies, topical ontologies, human/computer interfaces, computational linguistics, knowledge representation, information retrieval, and information-user studies. These researchers have worked with secondary literature, but most have not worked with primary documents. They can offer decades of experience in research design and methods in these areas.

Web interface developers and e-commerce companies. These commercial organizations need to know how people look for and find information, and what make good interfaces. Along with academic researchers, they have substantial

experience with web interface issues, as well as research and/or data-analysis results that may indicate practices to pursue and to avoid.

Publishers of both Web and print collections have an interest in intellectual access to electronically published text. In addition, they have test beds of data with which researchers might work.

Professional associations such as the Society of American Archivists, the American Library Association, American Society of Indexers, Association for Documentary Editing, the National Association of Social Studies Teachers, and many others include members with reason to care about these issues. Such societies can advocate for and endorse various research projects, as well as provide a forum in which they can be discussed.

Funding agencies interested in a wide variety of projects. They could include:

- federal granting agencies that sponsor research in computer applications, text understanding, digital libraries, educational research, and support for or use of cultural heritage repositories and their holdings.
- state granting agencies that sponsor the issues identified above at the local level and which may have local issues that relate to these questions.
- private foundations of all sizes and relevant interests.
- private industry such as computer manufactures, e-commerce companies, browser companies, and commercial publishers of electronic text, who have established research facilities and/or an interest in a better-functioning Web.

Conclusion

As increasing numbers of primary historical documents appear on the Web, publishers of those documents will need ways to provide intellectual access to the contents so the documents may be discovered and used. Today's search engines have limitations, although the advent of the Web has inspired many different types of researchers to increase the amount of work going into improving them. Most researchers, however, work with secondary materials which rarely present the severe problems of missing and ambiguous information that primary documents contain. Given that today's search engines can identify much that is in primary materials, publishers need guidelines on what technology can do, what requires human intervention, and how the two can be woven into frameworks for providing intellectual access to electronically published historical documents comparable to or better than that provided by a back-of-the-book index.

At a three-day meeting in Burlington, Vermont, experts in experimental electronic publishing, library and information science research, documentary editing, and computer science research identified (1) user studies to determine the needs and reactions of the audience(s), (2) implications for change

in publication management, and (3) technological approaches to access to information. These three general areas of research can contribute to our understanding of how to construct and improve intellectual access to historical documents on the Web.

Within these areas, the group identified a series of specific research issues which they painted in broad strokes so individuals and institutions can build research projects that serve the needs of the documentary editing community and others engaged in publishing historical documents or creating on-line cultural heritage exhibits. These areas of research can provide significant advances in the development of efficient and effective publication of historical documents, and may eventually lead to editorial guidelines to ensure full intellectual access for users. They may also lead to more systematic and interoperable standards for intellectual access across cultural heritage institutions. Without those standards, content providers and publishers will waste much of the time, money, and effort that goes into that sort of publication, while denying potential users access to the very materials they seek.

The alliance of archivists, librarians, documentary editors, and specialists in the information sciences seems to offer some of the most immediate benefits, both in terms of improving information retrieval and improved efficiency in preparing material for publication. As one participant later put it, "the most important practical outcome of our meeting in Burlington was the mutual realization of what [the editors of] multi-volume documentary editions and researchers in the field of computer-based conceptual indexing have to offer each other. . . . It's an opportunity left 'missed' too long and a collaboration that should begin as soon as possible and continue for decades to come."[21]

[21] Mary-Jo Kline, "Eight Steps, Not Twelve." Private e-mail to Elizabeth Dow, May 3, 2000.