CETH Workshop on Documenting Electronic Texts

May 16 – 18, 1994 — Radisson Hotel, Somerset, NJ

Lisa R. Horowitz

Technical Report No. 2

Center for Electronic Texts in the Humanities

Center for Electronic Texts in the Humanities
Rutgers and Princeton Universities
169 College Avenue
New Brunswick, NJ 08903
U.S.A.
Phone: +1 (908) 932-1384
Fax: +1 (908) 932-1386
E-mail: ceth@zodiac or ceth@zodiac.rutgers.edu

ACKNOWLEDGMENTS: Thank you to those who contributed notes for this report: Patricia Vanderberg, Annelies Hoogcarspel, Laurel Jizba, Douglas Kincade, Carol Mandel, Alan Morrison, and Brian Schottlaender. Thank you to the speakers who reviewed the reports of their talks: Randall Barry, Dominic Dunlop, Edward Gaynor, Richard Giordano, Clifford Lynch, Daniel Pitti, John Price-Wilkin, Allen Renear, and Michael Sperberg-McQueen, and to Peter Graham, Christine Bohlen, and Susan Hockey for additional proofreading. Finally, a very big thank you to Annelies Hoogcarspel who organized and coordinated the entire workshop.

The Center for Electronic Texts in the Humanities has received support from the National Endowment for the Humanities (an independent federal agency), the Andrew W. Mellon Foundation, the New Jersey Committee for the Humanities, and the Booth Ferris Foundation.

Copyright © 1994 Center for Electronic Texts in the Humanities

Table of Contents

1. Introduction 3
   1.1 Background Information 3
   1.2 Purpose 4

2. Overviews 5
   2.1 Needs of Humanities Scholars 5
   2.2 Introduction to SGML and TEI 7
   2.3 Networked Electronic Resources 14
   2.4 MARC 18

3. Practical Examples of SGML and TEI 21
   3.1 British National Corpus 21
   3.2 Berkeley Finding Aids Project 23
   3.3 TEI Header and Cataloging 25

4. Discussion Break-Out Groups 27

5. Conclusions and Recommendations 36

Appendix A: Program 39

Appendix B: List of Participants 41

Appendix C: Questions for Discussion 44

Appendix D: For Further Information 45

1. Introduction

In May 1994, the Center for Electronic Texts in the Humanities (CETH) sponsored an invitational workshop on documenting electronic primary source materials in the humanities. The goal of the workshop was to work toward a clearer understanding of the relationship between the TEI header, the MARC record, and the current international cataloging rules, with an objective of establishing how far they meet the needs of scholars, librarians, publishers and software developers who work with these materials.

This technical report is a summary of the overviews, discussion group sessions, and final recommendations and conclusions of that workshop, placed within the context of the goals of the workshop initially determined by CETH. Tapes of the workshop, along with some speakers' notes and hand-outs, the discussion group reports, and notes from several participants were consulted to create a comprehensive look at the workshop. CETH hopes that the report will generate interest and discussion in the long-term future of documenting electronic texts. The issues of using existing standards, creating new standards, and conferring between the scholarly community and the documentation community play a large role in this future, especially when seen against the frenetic enlarging of the Internet and the increasing number of non-standardized works in the field.

1.1 Background Information

Although the TEI and MARC are described in more depth later in this report, it is important to have some background information before beginning. The Text Encoding Initiative (TEI) is a major international project to develop and disseminate guidelines for the interchange of machine-readable texts among researchers in the humanities, and to make recommendations for the encoding of new texts. It is sponsored jointly by the Association for Computers and the Humanities (ACH), the Association for Computational Linguistics (ACL), and the Association for Literary and Linguistic Computing (ALLC).[1] The guidelines were formally published as *Guidelines for Electronic Text Encoding and Interchange* (referred to as TEI P3) in June 1994.

In these guidelines, a structure is proposed for an electronic text "header," somewhat equivalent to a printed book's title page. The header contains the traditional elements of a title page, such as title, author, and publisher. However, it also includes more extensive information specific to the electronic texts used by humanities scholars, such as explanations of the text encoding (e.g., what features were marked up: proper names, abbreviations, quotations, foreign words, bibliographic references, editorial comments, etc.); its source, if it was transcribed; and a revision history (who did what to the file when). This header can alleviate problems caused by the lack of documentation so common in informally-created (and even some formally-created) electronic texts.

[1]The preceding information was taken directly from Hockey, Susan, (1993), "The ACH-ACL-ALLC Text Encoding Initiative: An Overview," p. 1. Available at the listserv TEI-L@uicvm.uic.edu as TEIJ16.Memo.

MARC (MAchine-Readable Cataloging) is a definition of a structure for formatting data, originally designed as a standard for the bibliographic information found in library catalogs, although its uses have multiplied. In each record, MARC defines fields and subfields which represent specific kinds of information and which require certain syntax. Software for online library catalogs is based on MARC record formats. The software will generate a screen, based on each field's syntax and content, which is easy for users to read and understand.

1.2 Purpose

Originally, a workshop was proposed that would address issues surrounding the cataloging of electronic texts. As plans developed, the purpose of the workshop was clarified and broadened to include the convergence and/or divergence of MARC and TEI, because it became clear to CETH that the goals of these two electronic standards were similar, but the extent of the relationship between them was unclear. It was even possible to envision a single electronic file that would represent all the bibliographic information contained in the TEI header while also functioning as an access point, the way the MARC record currently does. Other issues which CETH considered important to both catalogers and users of the TEI guidelines were the question of what represents a new edition of an electronic text and what are the requirements for new electronic material. It was hoped that workshop participants would confront these issues and examine the implications for the computing, publishing, humanities, and library communities.

In preparation for the program, participants were asked to read a number of materials introducing the subjects under discussion. To illustrate the cataloging issues, the newly printed *Guidelines for Cataloging Monographic Electronic Texts at the Center for Electronic Texts in the Humanities* (informally called the CETH cataloging guidelines) and a booklet explaining USMARC formats (the MARC standard used in the United States) were included. To give background on TEI and SGML, three chapters from TEI P3 were included which explained SGML (Standard Generalized Mark-up Language, the mark-up language on which the TEI guidelines are based) and the TEI header. Additionally, the TEI header of an electronic text was included, with a related MARC record and a catalog record as might be viewed by a library patron.

The official program, a list of participants, a list of the questions posed to the participants, and a list of references for further information are included in the appendices at the back of this report.

2. Overviews

The one-and-a-half day program combined a great deal of new information with much discussion. For the first part of the workshop, experts knowledgeable about SGML, TEI and MARC presented overviews to ensure that all workshop participants, most of whom were expert in one or two of these fields, had a background in all three. The overviews began with an introduction on the needs of humanities scholars.

2.1 Needs of Humanities Scholars

Dr. Allen Renear is the director of the Scholarly Technology Group at Brown University and co-director of the Brown Women Writers Project, a database of pre-Victorian writing by women, encoded in SGML. Dr. Renear has an A.B. in philosophy from Bowdoin College, and an M.A. and Ph.D. from Brown. In 1983, after teaching philosophy for several years, he joined the staff of Brown's Computing and Information Services. He represents the American Philosophical Association on the Advisory Board of the TEI, and has written a number of papers on text encoding and humanities computing.

In the context of the workshop, one might expect this talk to encompass technological responses to humanities scholars' needs. In the talk, however, Renear went beyond the simple question of "what do humanists need" to what he referred to as the "meta-question": whether humanists (or anyone) actually *know* what they need (in regard to electronic texts or anything else). Using an analogy to psychoanalysis (what do women want? asked Freud), he explained what service specialists in any profession understand — that what a user requests is often not what he or she actually needs. Users often do not fully understand their own wants, and the service specialist, whether a systems analyst or a reference librarian, must work with users to discover these "buried" needs. The service specialist, standing at the point of contact between the support structure and the scholar, must interpret the expressed wants into actual needs and find ways to support those needs. Meanwhile, service specialists must also deal with the fact that scholars resent the implication that they do not know precisely what they need.

Renear's solution is a delicate balance that he terms "deferential aggression." Service specialists cannot in any way intimate that they do not trust a scholar's ability to evaluate his or her own problems. The scholar must be the ultimate authority. Therefore the service specialist must exhibit deference. However, at the same time, the service specialist must be aggressive in his or her attempt to get to the hidden need through patience and non-threatening questioning.

To illustrate this problem and solution, Renear gave an example of a humanist who was initially emphatic about what she wanted yet was very pleased with what she got, despite its difference from her original request. This folklorist requested a page-layout program for publishing exactly 99 nineteenth-century Romanian love incantations. What was discovered through patient discussions with her was that she actually needed a database to manipulate and analyze them.

Meanwhile, her assertion that she would only need room for 99 spells was based on a Romanian superstition, not a factual count; therefore Renear's persuasion to allow for 1000 spells (not just 100) proved invaluable, as illustrated by the 800 spells now contained in the database.

Renear also made an analogy to Marxist critical theory. The problems humanists have in identifying their interests and needs are not only based in their own individual ignorance, but are part of larger social contexts and ideologies. Therefore, he explained, service specialists must go beyond the individual encounter with the humanist to construct a larger infrastructure. Marxists say that we must not only understand the world but *change* it. Service specialists can do this by considering and developing those possibilities which scholars cannot even conceive because of their social context.

How can we place Renear's talk within the context of the workshop? An understanding of how to deal with the hidden needs of humanists will help us to drive the technology in the support system. What currently exists, such as non-standardized electronic texts or the MARC record, may be a simple solution to a want expressed by a scholar. However, what might truly be needed might not yet exist. Service specialists should consider these hidden needs of humanists, and think of the greater possibilities of technology in order to offer more. That could incite users to dig more deeply into their requests to determine what will really solve their problems. Understanding that scholars do not really know what they need and therefore cannot necessarily find answers in what already exists is the first step toward better technological solutions for the humanist.

2.2 Introduction to SGML and TEI

2.2.1 SGML

Dr. C. M. Sperberg-McQueen, the editor-in-chief of the TEI, is currently senior research programmer at the academic computing center of the University of Illinois at Chicago, where he has been since 1987. He is most recently involved with preparations for the university's information arcade, but has also supervised word-processing consulting and headed the database and campus-wide information systems group. He holds a Ph.D. in comparative literature from Stanford University, where he specialized in Germanic medieval literature.

Sperberg-McQueen gave an overview of SGML (Standard Generalized Mark-up Language), the mark-up language upon which the TEI guidelines are based. He explained first that SGML is a meta-language of mark-up languages, meaning that it defines what a mark-up language should be. This means SGML provides methods of specifying what mark-up is allowed and where, what mark-up is required and where, and how to tell mark-up from content.

2.2.1.1 Design Requirements

Sperberg-McQueen began by explaining the design features required in a mark-up language for use in encoding electronic texts in the humanities. First, an appropriate mark-up language must be able to represent all the characters of a text. Although the basic Roman alphabet and 10 numeric digits are easily encompassed by any mark-up language, some have problems with diacritics or non-Roman alphabets. For encoding texts in the humanities, a mark-up language must be able to mark the structure of the text, showing how it divides into paragraphs, lines, footnotes or other structural divisions. It must also be able to reduce the text to a linear order so that a computer can read it. This means embedding text such as footnotes and marginalia into the main body of the text without losing their structural integrity. A mark-up language must be able to represent extra-textual or non-textual matter, such as an analysis of the text or the images connected with it. It must have a clear method for distinguishing text from mark-up. Finally, because different kinds of computer applications require different kinds of information, a mark-up language must be both application- and system-independent.

SGML was found to be a highly suitable language for encoding electronic texts in the humanities because its basic features satisfy the above-mentioned requirements. SGML has other useful features as well. SGML is descriptive, not procedural. This means that SGML mark-up describes the text, but does not give commands about it. For example, a typesetting language might indicate "go to the top of the next page," while SGML would indicate "there is a page break here." SGML is not appearance-oriented; it would indicate "this is a stage direction" instead of "put this in italics." SGML is portable and system-independent, which is ideal for interchange between different systems and applications. Its definition is an international standard (ISO 8879), which

means that it is not owned by any corporation; users of SGML therefore cannot be locked into a particular system or application. It also has wide support because it is an international standard.

SGML requires a document type definition (DTD) (described more elaborately below), a document grammar which allows for data validation. This means that a program can mechanically ensure that a text is correctly marked up. For example, a DTD may state that titles only contain data. When parsing the text, if a title does not end before a paragraph begins, the parser will note an error based on the definition of a title found in the DTD. The DTD does not say what is wrong with a structure; only what is right. Therefore a DTD's creator does not need to envision all possible problems but instead simply states what is allowed. For example, the creator does not need to say that a chapter cannot start in a footnote; the DTD must merely say what *is* allowed in a footnote.

In SGML, text can be divided hierarchically into nested elements, and multiple hierarchies are possible, although not easily achieved. Any element can carry an attribute (a further description of the element; e.g., poems in an anthology might each be attributed to individual encoders), and there are entity references (simple string substitution mechanisms; e.g., you can declare TEI to mean Text Encoding Initiative without spelling it out each time). Finally, SGML uses consistent delimiters with very few special characters. This is all explained in more detail below.

2.2.1.2 Basic Concepts and Notation

Next, Sperberg-McQueen explained SGML's notation so the participants would be able to recognize it.[2] In SGML everything is delimited and everything is declared. Elements, such as poems, lines, stanzas, chapters or footnotes, are delimited by start and end tags. Tags are delimited by angle brackets (thus clearly separating mark-up from content), attributes by quotation marks, and entity references by a preceding ampersand and a succeeding semicolon. He gave a simple example illustrating these:

<quote lang='fra'> L'état, c'est moi!</quote>

In this example, "L'état, c'est moi" is the content of the element. The tags <quote> and </quote> mark the beginning and ending of the element. The attribute, "lang='fra'", explains that this quote is in French. "L'état" includes the entity reference "eacute" surrounded by the markers &…; indicating é.

The document grammar or DTD mentioned above is where everything is declared. Every element, attribute, entity, and even non-SGML notations such as images or sound, are declared and defined in the DTD. For example, an element such as "poem" or "anthology" is listed. Next to it, a

[2]Sperberg-McQueen noted that this section might be too technical for beginners; the following explanations are summaries of the basic points only.

symbol defines whether or not beginning or end tags may be omitted. This allows for decreased mark-up when possible. (For example, in a poem there is no need for an end tag, indicating the end of a line, because the start of a new line implies the end of the previous line.) Next to the tag omissability symbol, each declaration includes what other elements that element can contain. A poem, for example, might contain a title and stanzas. An example of a few lines of a DTD (in this case, not the TEI DTD) might be:

```
<!ELEMENT anthology - - (poem+) >
<!ELEMENT poem - - (title?, stanza+) >
<!ELEMENT title - O (#PCDATA) >
```

This says that an anthology and a poem require both start and end tags, but a title does not require an end tag (as designated by the O in place of the -). A poem contains one or no title (declared by the ?) and one or more stanzas (seen in the +). A title contains data (seen in the identifier PCDATA).

2.2.1.3 The Importance of SGML

Basically this was a simple introduction to SGML mark-up and document type definitions. It was hoped that the participants would be able to recognize, if not immediately comprehend, SGML mark-up. However, more important was the desire to explain why many humanities scholars find SGML and TEI crucial. Sperberg-McQueen finished his talk by explaining this importance. He emphasized certain benefits of SGML in particular:

- SGML makes possible a clean model of the text, allowing descriptive mark-up.
- Its independence from systems and applications makes it very useful for information-sharing; for example, it is ideal for scholars sharing authoritative electronic textual editions across different platforms.
- Because the mark-up reflects the *structure* of the document, it allows intelligent retrieval and document management.
- SGML can be used for complex documents, multimedia and hypertext because it allows for the incorporation of other, non-SGML notations.

He stressed its applicability to scholarly endeavors. Under TEI, mark-up can range from very light to very heavy. It supports many different kinds of analyses, even conflicting analyses. It can work with different kinds of textual requirements such as physical, rhetorical, and linguistic. It allows the many layers of information to be screened from display as needed, so that orthographic information, for example, can be visible at one time while the parts-of-speech are visible at another time, yet both of these kinds of information are always available.

Finally, Sperberg-McQueen reiterated that SGML allows for network-safe data exchange because it is device- and application-independent, as well as presentation-independent. Additionally, the

structure provided by SGML allows the transmission of structurally complete fragments of the text, rather than requiring the transmission of the entire text. This is easier on a network.

Sperberg-McQueen made it clear that SGML was a strongly supported way to encode electronic texts in the humanities. His explanations allowed those not familiar with encoding and textual analysis to get some background on the subject and to understand how standardized text encoding can aid in scholarly work on electronic texts. Also, standardized text encoding will mean that projects can be archived and still retrieved later by newer applications which are based on TEI guidelines. This talk laid the groundwork for the more specific discussion of the TEI header.

2.2.2 TEI Header

Richard Giordano is a member of the TEI's Text Documentation Committee, where he contributed to the TEI header. He teaches in the Department of Computer Science at the University of Manchester in England. He holds a Ph.D. in American history and an M.S. in library and information science, both from Columbia University. He has worked as a systems analyst at Columbia, a humanities computing consultant at Princeton, and a research scientist at Johns Hopkins University Medical School.

2.2.2.1 Background

Giordano began his talk by explaining how the header developed and who was involved in its creation. The Text Documentation Committee, who created the header, was made up of people from many professions with experience in cataloging, collecting and describing electronic texts. It included librarians, computer scientists and archivists, and publishing specialists, as well as representatives from various disciplines, including linguistics and natural language processing. The header's design reflected the TEI in general, because the structure of the TEI header was based on existing documentation standards, including AACR2 (*Anglo-American Cataloguing Rules,* 2nd ed., 1988 revision) and ISBD(G) [International Standards Bibliographic Description (General)]. The committee's work was also informed by other TEI work groups, especially the work group on spoken language.

2.2.2.2 Structure

The header is documentation which is "prefixed" to every TEI-conformant text. It is made up of four sections, only one of which is mandatory: the file description <fileDesc>, the encoding description <encodingDesc>, the text profile <profileDesc>, and the revision history <revisionDesc>. These are each explained in detail below.

The file description <fileDesc> is the electronic equivalent of a title page. It contains a full bibliographic description of an electronic file, and is the part of the header which is most closely aligned with the MARC record. It is the only required section of the header and is intended to be

used for the creation of a standard library catalog record. The elements of the file description closely resemble the MARC record. There are three mandatory and four optional elements. Giordano pointed out that all the elements are not mandatory, because the committee felt that if there were too many requirements, people would not comply at all. By requiring only some elements, the committee hoped that creators would be encouraged to include at least those. The required elements are a title statement (which includes the title of the electronic text and its author), a publication statement (including publisher and date of publication), and a source description (a description of the source from which the text was transcribed). The optional elements are the edition statement, the extent, a series statement, and a notes statement. The elements of the file description allowed for inclusion of all those responsible for the intellectual content of the work, such as author(s), sponsor(s), funder(s) (if different from sponsor), and principal investigator(s)[3].

In the file description section, as well as the others explained below, there is usually no required syntax within the elements, although structure is recommended. Prose may be used in place of structured elements, if preferred. For example, a creator could write[4]

<titleStmt>Two stories by Edgar Allen Poe: electronic version, by Edgar Allen Poe, compiled by James D. Benson</titleStmt>

or could use a more structured format:

```
<titleStmt>
   <title>Two stories by Edgar Allen Poe: electronic version</title>
   <author>Edgar Allen Poe</author>
   <respStmt><resp>compiled by</resp>
   <name>James D. Benson</name></respStmt>
</titleStmt>
```

The header was intended to be used by both catalogers and non-catalogers alike, so emphasis was placed on content, not on syntax or strict rules. The desire was that creators should supply information. The committee did not expect the creators to act like librarians, nor did they want to impose rules on the library community. By allowing some flexibility in the header, the committee allowed for the provision of necessary information without too many demands, with the expectation that catalogers would still use their own methods to make everything consistent.

---

[3]The term “principal investigator” engendered quite a bit of discussion because the principal investigator in the TEI guidelines is the one who adds intellectual content to the electronic work, such as an editor of an electronic *Moby Dick.* The argument was that in the hard and soft sciences the principal investigator is the one who controls the intellectual content of a study, while the TEI’s definition sounded more like an editor of the work. The argument was left unresolved.

[4]Examples are taken from C.M. Sperberg-McQueen & Lou Burnard, eds., (1994), *Guidelines for Electronic Text Encoding and Interchange,* Chicago: Text Encoding Initiative, p. 96.

The encoding description <encodingDesc> describes the relationship between the encoded text and its source. Giordano likened the encoding description to the codebooks used in the social sciences to explain data. He gave several examples of what might be usefully included in this section, aside from simply explaining how certain textual features, such as proper nouns or foreign words, were marked: If errors in transcription were controlled with a certain spell-checker, that might be important to mention in the encoding description because the spell-checker might refer to American English while the source was in British English. Or, if a work were encoded by a number of different graduate students, the encoding description might identify each student with an attribute which would then be used wherever that student encoded text. The encoding description is optional, but it is highly recommended. An encoding description with clearly defined encoding practices is invaluable to researchers but is often not included in a standard MARC record.[5]

The profile description <profileDesc> is used mostly for transcriptions of speech, and it was this area which was most influenced by people whose expertise is not traditionally in documentation, particularly the spoken language working group. It is an optional section which provides information on the language, situation, text, participants and settings of spoken texts. For example, the speech of an excited man speaking with a New York City accent can be isolated and categorized in the profile description, and used to support both lexical and linguistic analyses. Standard bibliographic practices do not generally document this kind of information. The profile description also allows a creation date, where the date of the conversation might be listed, as opposed to the date of the transcription or the date of publication of the electronic file.

The final section in the header is the revision history <revisionDesc>. This section lists any changes made to the electronic file, encompassing who made each change, what was changed, and when the change was made. The people included in this section might be data entry people, encoders, editors, or someone who made changes after publication. It does not need to include all of these people, but it can, depending on the individual creator. The revision history is a vital part of an electronic text as it may reflect version/edition control issues. This section is also optional.

Giordano reiterated that the four sections of a TEI header are not intended to imitate catalog records, but are intended to inform them. Few of the sections are mandatory and input can be in prose form, although the structured elements and optional sections are desirable because they make the text easier to process and more informative to document that text. Giordano noted that use of a full tag set could make the header larger than the text being documented. The size of the header really reflects the nature of the project and the amount of documentation desired, or even simply the funding and human resources available. A header which is larger than the text is not inappropriate when the situation warrants it. The header can range from skeletal or extensive, again depending on the goals of the project.

[5]The CETH cataloging guidelines do describe how to include some encoding features in the MARC record.

2.2.2.3 Comparison of TEI Header to the MARC Record

Giordano's comparison of the TEI header to the MARC record was from the perspective of a librarian. He stressed that the header was not intended to be used as a catalog record. There is a fundamental difference between the two which implies that there should always be two different electronic locations of the information, one which is allowed to be full but loosely structured and a second which is more structured but possibly more sparse. He described MARC as a telecommunications format for a catalog card, containing bibliographic information. The TEI header was intended to provide full bibliographic information, with internal pointers to support both machine and human analysis. MARC is much more granular while the TEI header is much more functionally oriented.

Mapping from the MARC record to the TEI header may possibly be done through one of two ways, Giordano explained. First, the profile description, the least bibliographic element of the header, can be placed into a 590 structured note field. Second, a sort of "codebook" could be created which contains all the descriptions aside from the file description; then a pointer from MARC could be made to that codebook. The codebook could then point to relevant texts or sections of it. Overall, Giordano emphasized his dislike of "shoehorning" non-bibliographic information, such as that included in the profile description, into a MARC record.

2.3 Networked Electronic Resources

Dr. Clifford Lynch is the director of the University of California Office of the President Division of Library Automation. He oversees the MELVYL public access information system, and is involved in a variety of research and development projects. He is well known for his writings and talks on library automation and information science. He is on the board of NISO (National Information Standards Organization) and is leader of the Architectures and Standards Working Group of the Coalition for Networked Information. He came to the UC system about fifteen years ago from New York University and holds a Ph.D. in computer science from UC-Berkeley.

2.3.1 Context

Despite Lynch's claim that he was having trouble fitting in his topic with the previous talks, he succeeded in describing some of the very complex issues involved with making electronic texts in the humanities available over the network. The issues that he discussed related to questions of organization and access, as well as standards, all of which come into play under the umbrella of documenting electronic texts.

Lynch's talk focused on what is happening over networks with all kinds of electronic objects, not just electronic texts, including graphics, audio, and home pages on the World Wide Web. He compared the scrupulous care that is used to describe a document using the TEI standards to the "Internet mess," which has been encountered by anyone trying to find a vague item on the Internet. He explained that abstracting and indexing databases often describe material without clearly identifying in which edition of a newspaper or periodical the material appeared (a growing problem given the number of regional editions now being published for newspapers and magazines). If the article itself is available in electronic format, it often omits charts, images, and tables, frequently without a clear indication of what intellectual content is actually missing. This creates a "poor provenance for reconstruction" or use.

2.3.2 Underlying Issues

Providing access to items over the Internet currently does not involve the content of the item so much as it means citing and finding objects; it deals very little with bibliographic description. One of the core problems encountered in organizational access of electronic items is edition and version control. In paper volumes, a new edition usually means changes to intellectual content. Occasionally, though, a publisher will designate a new edition when the changes are merely in layout and pagination. In the digital world, the issues are more complex. A new edition could still mean a change in intellectual content or a change in formatting. However, it is very easy to compare two electronic versions and differentiate them at the level of bit equivalence. This could mean a single difference between two documents, such as one word spelled differently (perhaps just corrected) or even typographic or font changes. Do these changes merit consideration as a new edition? Are they significant?

This idea of version control, very popular with a number of people in the computer science community, is simple to implement but eliminates the concept of intellectual similarity. Lynch advocated the implementation of some way of grouping versions which are not bit equivalent but which do not differ in intellectual content. This would allow users to zoom in and out when viewing the results of a search, looking at citations at different levels of granularity. For example, at one extreme, a user might see only major intellectual editions of works of Shakespeare; at an intermediate level, he or she might see an enumeration of different editions (in the print publishing sense) of each play; and at the other extreme, the user might find listings of very large numbers of texts that perhaps only differ in mark-up, source of transcription, or representation in electronic format.

2.3.3 Identifiers on the Internet

In the network world, the focus in organization and access is to the digital object, not to the intellectual content. This has brought about three succeeding forms of identifiers for the network: Uniform Resource Locators (URLs), Uniform Resource Names (URNs), and Uniform Resource Characteristics[6] (URCs).

URLs were brought into very broad usage through the World Wide Web and its interfaces. A URL is an identifier of a network service and a package of parameters to be passed to that service. The result is a digital object or a connection to a network service, depending on the type of URL. It is a link to a particular place on the network, and can be used in accessing different kinds of online resources, such as telnetting, ftp'ing, and browsing in a Z39 database. A URL is a pointer or path to an object. It is not reliable in the long-term because it is so easy to move a file or change a host name, thus invalidating the former URL. Given a URL, a user may find that the object it points to no longer exists, and this is usually without warning. Additionally, it is possible to have multiple URLs for the same item; for example, if the same item were in multiple places or if you could get to an item via telnet, ftp and gopher, there would be a number of URLs for it.

A URN, on the other hand, is intended to have some sort of permanence as a persistent name. A URN is an alphanumeric "name" which could be used in a scholarly work to identify a reference. It is good in the long-term, even if the actual item is no longer available. An example of models for URNs might be an ISBN (International Standard Book Number) or a patent number. Even if you can no longer buy a certain book, you can usually identify it by its ISBN. The nature of a URN requires some sort of resolvers which would map from a URN to its URLs. Likely resolvers already exist, such as online catalogs, union catalogs, or even publishers' catalogs, where the URN could be connected with a particular item while the URLs could be updated as different or new ones became known.

---

[6]The form "URC" is an established one, but Lynch pointed out that what the C represents has varied; "citation" has also been used occasionally. For further explanations of URLs, URCs, and URNs, see the works referenced in Appendix D.

URNs are actually made up of two parts. The first is a naming authority identifier. These will be established by an "intergalactic" authority that will register individuals or organizations who wish to assign "names" (URN identifiers). These individuals or organizations will serve as naming authorities. They will assign URNs and decide when a file is a new edition. Depending on the naming authority and why it is assigning URN identifiers, there will be varying practices in deciding when something is a new edition of a work. This could imply some chaos unless standards are in place for defining an edition. For example, if the U.S. government became a naming authority, it might choose to give its brochure *How To Get a Social Security Card* only a single URN regardless of the version. Previous versions of the brochure would then not be accessible via that URN. An archive, on the other hand, might assign different URNs to each edition of the same brochure. Some naming authorities could choose to implement a system where new names would be assigned based on bit equivalence, thus creating many works of intellectual similarity without a clear indication of when the intellectual content has changed.

The URL is currently a stable syntax which is close to being finalized. It is open-ended, though, and can accommodate new protocols as they are available. So although ftp, telnet, WWW, and gopher are available today, new URLs can be devised as new systems become available. The URN standard is less mature, and the syntax of URNs is still being discussed. While URLs are immediately useful as soon as they are created, URNs assume the existence of a large infrastructure with naming authorities and resolvers which map URNs to URLs. This infrastructure is not defined yet. When this has been resolved, URNs will act as a bridge between the bibliographic record at the item or collection level and the source material. Systems are being envisioned where a user might retrieve an electronic item by clicking on a bibliographic citation obtained from an online catalog. The system would automatically resolve a URN in the bibliographic record to a URL, and then evaluate the URL to retrieve the electronic item from the Internet.

2.3.4 The Future of Internet Identifiers

URCs are much more complex, and there is as yet no agreement on the precise scope of what a URC should contain. URCs were originally intended to be human-readable and to be somewhat analogous to the citations that occur in print literature. A URC might be made up of a URN, the relevant URLs, and assorted meta-data such as the title, author, and publication date of a text in the humanities. Currently in this field, people outside of the library have been trying to define appropriate meta-data without recourse to the work that librarians have already done, and may be well-served to reference the MARC record and its history. It has also been recommended that other information needed for network-based retrieval systems — such as file size and format, which really do not relate to intellectual content — be included in URCs as well, bringing up the recurring issue of representation and format versus intellectual content. People investigating the creation of these URCs are mixing author and title information (i.e., intellectual content) with information about which software was used to create the text, such as Microsoft Word 5.0 or

PostScript (i.e., representation and formatting issues). A consensus must be determined defining what a URC should contain.

Lynch concluded his talk by returning to the TEI and SGML. SGML has the ability to link to the network via a filename and perhaps a URL. In the future, it will probably have ways to link through URNs and URCs. Lynch asked what role text archives will play in assigning URNs for electronic text. He also questioned what information about electronic texts might be useful in a URC; perhaps information contained in a TEI header.

2.4 MARC

Randall Barry is the Senior MARC Standards Specialist at the Library of Congress. He has worked at the Library of Congress since 1977 in several positions. Since 1987, he has been responsible for the USMARC format for bibliographic data, foreign MARC conversion and format, and character sets. He has also served as convener for the ISO working group on character sets since 1991. He has degrees in French and Russian from the University of Maryland, and an M.L.S. from Catholic University of America.

2.4.1 Structural Components

The purpose of Barry's talk was to give an overview of MARC so that non-librarians would be able to follow the workshop discussions. MARC was originally designed as a standard for the bibliographic information found in library catalogs, but is now used in various locations, including archives and museums. It is not a system but a structure. A MARC record consists of three components: the record structure, the content designation, and the data content.

The record structure is a de jure standard which is documented and applied as national (ANSI Z39.2) and international (ISO 2709) standards. It is a way to make all data in a record specifically identifiable, similar to SGML's marking up text to be structurally identifiable. The record structure is less complex if conceptualized as a string of numbers and letters controlled with end-of-field, end-of-record, and delimiter characters.

There are three parts to the structure: the leader, the directory and variable fields. The leader is the first 24 character positions of the record. These positions contain specific codes. For example, in a certain location in the 24 positions, an "a" means that the record was cataloged under AACR2 cataloging rules; another group of positions contains the number of characters in the entire record (as a back-up for the end-of-record character). The directory is a series of 12-character entries, one for each field. Each directory entry shows the tag for the field (a three-digit field identifier), the length of the field in number of characters, and the address of the first character of that field. This tells the program where to find each of the variable fields which follow, and defines what those fields are. This record structure explicitly identifies the data elements and codifies it so that it can be programmed for display.

The content designation, the second part of the MARC record, makes the data intelligible. Each content designator is made up of a three-digit numeric tag which has a specific meaning based on the MARC format used (see below). Often these tags are presented in conjunction with the data they refer to, such as in some cataloging software packages and some online catalogs, but in the record structure, the tags reside only in the directory. For display, these tags can be translated into words to make the record easy for the user to understand. The content designation also includes two numeric indicators of one digit each, for each variable field. These can be used to fine-tune the definition of a field. For example, the 100 field in the MARC format for bibliographic data

represents the primary author of a work. The indicator values 0 through 3 each represent the way that the author's surname is represented; John-Paul, Pope, for example, would get a 0 in this indicator position, indicating that he has no surname. Finally, the content designation includes subfield codes which further divide up a field into its component parts. These are usually represented by lowercase letters, and their syntax usually requires a particular order that is not necessarily alphabetic.

The data content in a MARC record is not defined by MARC. Cataloging rules dictate what information is included in a MARC record. Cataloging is affected by international standards, local or national laws, and policy statements. MARC then allows the implementation of these rules and laws through MARC formats created for different types of records.

2.4.2 MARC Formats

Barry compared MARC formats to SGML DTDs in that they are lists of valid content designators for specific record types just as SGML DTDs are lists of valid tags for specific document types. Each MARC format defines each character position in fixed fields, and the available tags, indicators, and subfield codes. The format also defines which fields can be repeated and which are mandatory.

Barry emphasized that the MARC formats which exist now represent 25 years of standards work. Over time, MARC has spread from country to country, changing somewhat in different implementations. The Library of Congress now publishes five different USMARC formats. USMARC is the implementation of the MARC structure used in the United States. The format which is used in cataloging library materials is *USMARC Format for Bibliographic Data.* Other formats are used for authority, holdings, classification, and community information. USMARC is the only record format used in the United States for the exchange of library data.

2.4.3 System Requirements

Systems which use MARC must meet certain requirements. They must be able to import and export MARC records with round-trip compatibility. This means that if an imported record is then exported back to its original system, it should be unaltered, including order of fields. This is an important requirement because a programmer ignorant of library cataloging rules might force subfields to sort alphabetically, but, as mentioned above, the standard order for subfields may not be alphabetical. The system should not, therefore, make these kinds of changes. A system using MARC records must allow for the creation, modification and deletion of MARC records, and must be able to handle the bibliographic character set.

2.4.4 MARC vs. SGML

Before his talk, Barry was introduced with the comment that people are rethinking whether MARC serves a logical purpose today, or whether perhaps it is time to replace it. Barry felt that MARC legitimately holds its own in today's electronic world, and is superior to SGML for its particular purposes. He pointed out that although SGML and MARC were both designed for exchange, retrieval and output products, SGML was designed for large quantities of text. MARC was designed for large numbers of small records. In addition and probably more important, MARC has particular strengths in its use in libraries. It has been used since 1968 in libraries, and because of this, there is a high level of conformance to MARC standards. Library systems around the world are designed around the MARC record structure, and the systems and software would all need to be redesigned to support any other structure. Barry's conclusion, alongside the three other speakers' own views, laid exactly the right groundwork for the later break-out discussions.

3. Practical Examples of SGML and TEI

Three short presentations illustrated actual use of TEI and SGML in projects in order to better familiarize participants with these tools.

3.1 British National Corpus

Dominic Dunlop is former task manager of the British National Corpus at Oxford University Computing Services. He was trained as an electrical engineer, and then moved into commercial computer programming, support and management. His first exposure to academic projects came three years ago when he began work on the British National Corpus project at Oxford.

The British National Corpus is a database of 100 million words of modern British English. It was established for commercial lexicographers for creating dictionaries, and for teachers and researchers of the English language. Eighty percent of the corpus is from published written sources, with another ten percent from unpublished sources such as internal memos, sermons, private correspondence and electronic mail. The remaining ten percent are transcriptions of spoken sources. Half of these were "context-governed"; specific contexts, such as a trade-union meeting, were chosen by the British National Corpus staff. The second half were conversations taken from a market research sample of people from representative social classes and regions who carried recorders until they filled ten cassettes. The entire corpus is marked for parts-of-speech by a parsing program.

The corpus is encoded in SGML under TEI guidelines. Because the corpus was begun in 1991, and many of the features in the current version of the TEI guidelines (TEI P3) were not yet in place, these new features could not be implemented in the mark-up of the texts of the corpus. Also, the amount of mark-up depended on time and resources, so that the mark-up is not necessarily as complete as would be desired. The TEI headers created for the corpus, however, were done later in the process and thus followed fairly complete TEI guidelines. Dunlop focused on the headers in his talk.

The headers were generated automatically from a relational database where information about the texts was stored. Two kinds of headers are used. The corpus has its own header, and then each individual text also carries a header. The corpus header contains as much as logically fits within it, while the text headers, because there were thousands of texts, were kept as "lean" as possible. This created a bibliographic problem in that text headers cannot stand alone. There are pointers from the text header to the corpus header, where there is fuller description or explanation of features which were duplicated across texts.

The corpus header includes a description of copyright and availability of the corpus, a bibliographic record for the corpus itself, and a description of the make-up of the corpus which includes how the texts were classified. Dunlop noted that the bibliographic record for the entire

corpus was challenging to create because there are no standards for what should be in a corpus bibliographic record. The corpus header also incorporates any information related to all the texts in the corpus. Even when information is connected to only two or more texts, it is included in the corpus header with pointers from the relevant text headers.

The text headers include specific copyright descriptions and a bibliographic record for that particular text. The bibliography includes the source text information as well as the electronic text information. Anything relating to the individual texts, such as odd editorial practices, is also included in the text headers. The text header contains an explanation of why that text was included and where it belongs within the "sample space" of the corpus.

Dunlop listed a number of problems encountered when incorporating TEI headers into the corpus. Problems which came from within included a problematic work flow for the production of the corpus, inconsistency in the use of bibliographic information, and lateness in defining final header structure and an overall DTD. External problems were more complex. Some of these had to do with insufficient or inappropriate definitions of elements within the TEI. For example, a source description is mandatory in header information, but how is the source of a corpus best defined? Other problems related to finding a location for information that seemed valuable yet did not fit into the TEI structure. For example, it is unclear where to include details surrounding the life of an author (birthplace, birth and death dates, regional accent, etc.), or the habits of the speakers (such as what newspapers they read, important details to a linguist). In a final comment, Dunlop remarked that because standards are so prevalent for published texts, there is a bias towards published works in the TEI. Despite the additional features added to the TEI header by the working group for spoken language, Dunlop and others at the British National Corpus project found that it was difficult to encode headers for spoken texts.

3.2 Berkeley Finding Aids Project

Daniel Pitti is Librarian for Advanced Technologies Projects at The Library at the University of California at Berkeley and the database design coordinator for the Berkeley Finding Aids Project. Prior to this, he was the authorities librarian at UC-Berkeley and was responsible for, among other duties, designing the authority control subsystem in the online catalog at UC-Berkeley.

The Berkeley Finding Aids Project is a Department of Education Title II-A project with two goals: to develop a prototype encoding standard for libraries, archives, and museum finding aids, and to develop a prototype database of encoded finding aids. The encoding is done in SGML, but is not TEI-conformant (although Pitti mentioned that he felt some pressure at this workshop to make it so). Finding aids are documents used to control and provide access to unpublished collections, usually of primary source materials. They have traditionally been printed documents, accessed through the card (or online) catalog. Archival finding aids provide a hierarchical analysis of a group of documents usually gathered by provenance, such as a collection of someone's papers. Within the collection, subcollections are noted, and often the finding aid will lead to individual items. The organization depends very much on the type of control and access desired, such as archival, museum, or bibliographic.

Using SGML's ability to link to non-SGML notation, electronic finding aids can do the same as the former printed ones, but better. They can link at the individual item level directly to a digital surrogate of the primary source material itself, while still giving access to that primary source. A digital surrogate gives anyone with access to the finding aid access to the content of the material itself, thus making the item available beyond a group of specialists or those who can physically come to a library. The use of digital surrogates could also possibly be used to preserve collections by reducing hands-on use. With these thoughts in mind, the people at the Berkeley Finding Aids Project have begun creating some electronic finding aids. In these finding aids, access to an item is top down. The top level is a MARC record of the collection. This record points to the SGML-based finding aid. That in turn points, via its own hierarchy of subcollections and lists, to digital surrogates as well as to the primary source material itself.

Pitti exhibited an example of a finding aid, part of a cooperative project with RLG and Stokes Imaging Services in Austin, Texas. This collection was a group of 1000 images from the late nineteenth to early twentieth century collected by photographer Roy Graves. The finding aid was shown using Electronic Book Technologies' DynaText, an SGML-aware browser, to create a top-down description of the collection. It opened up to a table of contents derived from the SGML mark-up of the finding aid. The table of contents was hyperlinked to different portions of the finding aid, giving random access to the aid. Within the table of contents, one can view just the top-level items, or one can delve into the deeper levels of the hierarchy. Pitti pointed out that the introduction to the finding aid contains the same information found in a MARC AMC collection-level record.

DynaText formats the electronic text for display in real-time using the SGML-based structure of the text. Exploiting the structure, DynaText dynamically reformats the text to fill the available screen size, wrapping the text if the window is made narrower. It is even possible to enlarge the font size dynamically. Because the SGML-based text is marked descriptively rather than procedurally, it is possible to produce both printed and online finding aids from the same source file. It would even be possible to produce Braille output.

One of the special features of this finding aid was how the numbering system of the inventory of the collection was displayed on the screen. The inventory was comprised of four series which were each comprised of subseries. This was reflected in a physical system of outer and inner containers. In a printed finding aid, often the outer container is assumed and therefore not listed next to each item. In the electronic finding aid, Pitti illustrated that the number of the outer container is visible in the table of contents, allowing users to know at a glance where an item is contained.

Pitti concluded with a comment on the platform. DynaText was shown using Windows on a PC. Currently, Electronic Book Technologies has browsers that work on X-Windows in UNIX, as well as in Macintosh and Windows formats for individual PCs. A variety of SGML browsers are available from other vendors as well. Pitti remarked that, using DynaText, Berkeley will make the prototype database of finding aids available over the Internet for users with X-Windows software.

3.3 TEI Header and Cataloging

John Price-Wilkin and Edward Gaynor work at the University of Virginia, Price-Wilkin as Systems Librarian for Information Services and Gaynor as the Head of Original Cataloging. Price-Wilkin has been at the University of Virginia since 1992. Prior to that, he served in several positions at the University of Michigan where, among other things, he implemented a full-text search and retrieval system providing online access to a wide range of electronic texts. He has an M.L.S. from the University of Tennessee at Knoxville, and an M.A. in English from the University of Virginia. Gaynor has been at the University of Virginia since 1989. His M.L.S. is from the University of South Carolina.

Price-Wilkin began the talk with some information about the electronic texts held at the university. Texts encoded in SGML almost exclusively, with a majority following TEI guidelines of one of the three drafts, have been made available campus-wide, and some are available over the Internet. The purpose of the electronic text collection is to support textual analysis with an open-text philosophy. At the time of the presentation, the collection was 5,000 items, over one gigabyte of raw data, provided on a network to those affiliated with the university. Some texts are also available publicly. The texts are accessible over a number of platforms. Price-Wilkin emphasized that it was critical to the library that access be over the network because, despite the existence of an electronic text center in the library, they wanted the texts available where people do their work.

There are currently three primary interfaces, all of which rely on Open Text Systems' PAT program, an SGML-aware search engine. The code for allowing access over the network is freely available. The interfaces are analytical, allowing, for example, queries of the structure and proximity queries (the types of queries that humanities scholars might pose). With the help of Mark Olsen (another participant at the workshop), Price-Wilkin has created a World Wide Web gateway to PAT, which allows for textual browsing, a function desirable to those who are less likely to use the analyzing features offered in other formats. This interface has buttons for querying. There is a unique URL for each text, even when the texts are within a collection. Although the richer, primarily TEI-conformant SGML is used to encode the texts, results of searches or browsing are filtered to HTML (HyperText Mark-up Language — an application of SGML used primarily for Mosaic, a World Wide Web interface) for displaying the texts.

Price-Wilkin explained that TEI headers are used to provide access to the materials via what he called "subsetting," which enables a user to look at a subset of texts formatted through various SGML tags, such as creation date, author, and classification characteristics, or to look at the header directly. Many of the texts come to the library in SGML, but not all, and much work is done to encode them. However, Price-Wilkin noted that there was a limited understanding of the bibliographic description required in the header on the part of the encoders at the electronic text center. This, he felt, provided a unique opportunity for catalogers to get involved in the process of encoding electronic texts.

Gaynor described the environment at the University of Virginia which led to the encoding of TEI headers by catalogers. The library had established a policy two years ago regarding its future goals in terms of what materials to provide access to, and how to provide access. When electronic texts were first acquired, the library determined that these fell within the guidelines and that it was well worth the effort to make electronic texts a priority. Although the catalog department was not offered additional resources to take care of this new priority, they were allowed to juggle their current priorities. Other cataloging was dropped from a professional to paraprofessional level, and when a professional position was vacated, the position was redefined as an electronic resources cataloger.

The decision to provide full access to an electronic text environment led logically to the request for catalogers to encode TEI headers, and the catalogers found the encoding of text headers a unique way to contribute. Gaynor underscored that the catalogers' skills and knowledge in arranging bibliographic information were an incredible asset to creating the headers, while to the catalogers, SGML was just another tag set. The encoders in the electronic text center give the catalogers a checklist (on paper) which includes the information that the encoders have about the source and file revisions. The encoders also include a minimal header at the top of the documents. The catalogers have seen from these documents that their own understanding of a bibliographic citation is critical and that their skills are strongly needed, giving them, said Gaynor, a boost. The citations on the checklist can be vague and unclear, and the catalogers need to do quite a bit of research to find, for example, the actual texts that were transcribed. The catalogers create complete file descriptions in addition to the less bibliographically-oriented encoding, profile and revision descriptions. The headers are then parsed using an Oxford Text Archive DTD to mechanically ensure that they are marked correctly. The header is created in a separate file which is then appended to the text file.

The MARC record is created from the information in the header. All header entries are formatted according to AACR2 and Library of Congress cataloging rules to save time later. Gaynor described the problems involved with this dual process of creating headers and then separate MARC records. Most important, it is incredibly labor intensive. The reasons for this relate to some of the existing mechanisms used for cataloging: For most cataloging, the University of Virginia contributes to OCLC on tape. For technical reasons, however, OCLC cannot accommodate the university's original computer file records when contributed on tape, so they must be entered directly on OCLC terminals. Therefore, there is a lot of rekeying and manipulating of data. Gaynor suggested a stronger compatibility among systems to make this process easier. Gaynor concluded with a recommendation that the TEI include standards for inputting AACR2 fields into the headers. This would allow for the sharing of data between systems and institutions, and would possibly prevent some of the double work which his catalog department is currently required to do.

4. Discussion Break-Out Groups

The participants were divided into four groups of approximately twelve people each. An attempt was made to include people with different perspectives in each group. Each group's moderator kept the discussion focused on a prescribed set of questions. All four groups discussed the same set of questions. The questions (see the complete list in Appendix C) covered topics such as the information needed to describe or access electronic texts, the special needs required by various formats such as images or sound as opposed to text, and dealing with different versions or editions of electronic materials. Break-out discussions lasted four hours, followed by a general discussion of an hour and a half. A reporter recorded his or her group's viewpoints for presenting at the final discussion. The following are summaries of the reporters' presentations.

4.1 Group 1

Moderator: Nancy John
Reporter: Carol Mandel

Participants

Judy Brugger	Peter Graham
Arianne Burnette	John Price-Wilkin
David Chesnutt	Sarah Thomas
Steven DeRose	Patricia Vanderberg
Marianne Gaunt	

The moderator began the group discussion with a reiteration of the purpose of the workshop: to discuss the relationship between the TEI header and the MARC record. This led to a discussion of merging the two electronic forms of bibliographic information in some way. The role of the header was considered from two angles: Does the header enrich the MARC record, or should the MARC record be placed in the header? Although the group agreed that the two different items were redundant in the material they contained, they determined that to merge the document with the access tool was not a good idea. It is unlikely that the data producer, normally the one to create the TEI header, would be capable of or have the desire to create a MARC authority record. The group felt that the TEI header, which could be used as a mechanism for browsing information contained in the text, should remain embedded in the text, while catalogers should retain authority control[7] via the catalog record.

[7] Authority control is "a technical process executed on a library catalog to provide structure. Uniqueness, standardization, and linkages are the foundation of authority control. … The operation involves research, the creation of standardized forms of access points, and linkages to variant forms. As a result of meticulous research, a name is unique, that is, distinguishable from other names that may be candidates for inclusion in the catalog at some later date." [D. H. Clack, (1990), *Authority Control: Principles, Applications, and Instructions,* Chicago: American Library Association, p. 1.] For example, authority control is what allows a user to use either Mark Twain or Samuel Clemens to find one of his works in a library's catalog, even if the title page of the work only lists one of these two names or any other variation.

The group defined the separate roles of the header and the catalog record. The header is used as the basis of the bibliographic record. It is an important point for version control and revision history of the text. The catalog record, on the other hand, is used for access to the text, and thus is the appropriate location for authority control. The value of the TEI header is in the description it provides, not in uniformity of style or names. The value of the catalog record lies in its use in information retrieval through the authority forms of names, as well as through subject access.

The model envisioned as the access mode of the (probably immediate) future puts the MARC record as the first point of access, linking to the header and to the text itself. The TEI header would be the next access point, containing very full bibliographic information, and could be used for browsing. How would this best be mechanized? The group proposed three possible scenarios: (a) An integrated approach between creators and catalogers, such as that used at the University of Virginia, (b) a layered approach where the cataloger creates the MARC record while someone else, most likely the creator and the publisher, produces the TEI header, or (c) a hybrid approach, where the header is produced by the creator, enhanced by catalogers, and automatically generated into a MARC record.

This last option gave rise to many recommendations. In order to achieve this hybrid approach, some changes should be implemented in the MARC record and in the TEI header. In the MARC record, it would be useful to have fields that announce TEI conformance and non-text formats. Clear version information should be included, and new access points should be (and are being) planned for electronic documents. The MARC record would not replace the TEI header's bibliographic description, and might even be reduced by the header, while the TEI header and document must be able to stand alone.

The TEI header could be improved for this envisioned future use as well. One recommendation emphasized that headers must be created to allow readers to browse easily or to move to a certain section of the document. The header should also summarize the non-text formats linked to the document in a more extensive prose form than a DTD would. This would allow illustration information to be translated into the catalog record, and it would give users information about what is needed to view linked images or to connect to other parts of the document. The group suggested that standards and more precision be defined for the revision history section, which could then be better mapped to MARC for generating the catalog record.

The group also gave other more general recommendations. They felt that the header creators need good guidelines and training to create quality headers. There should also be guidelines and automated support for mapping from the TEI header to the MARC record, to make it possible to generate at least preliminary catalog records automatically. This would greatly accelerate the cataloging of the texts.

The group spent some time elaborating on the version/edition control issue. They felt that more work must be done to clarify what a new edition is: Specifically, they determined that different

mark-ups and editorial revisions constitute different editions and must be cataloged as such. The group felt that naming and authentication systems should be established (URNs plus a digital signature) not only to trace and find items, but to identify and verify electronic versions. A different URN for each version would enable machine authentication of versions, as well as better recording and identifying of versions. This name or number would map easily to the catalog record.

Finally, the group described the roles various members of the scholarly community would play to make this system work. Text creators, publishers, librarians, and system designers all have important jobs in the documentation of headers. The creators of the electronic texts can provide the most complete information in the headers, and they should be motivated to do so. Publishers, who are accustomed to standardizing bibliographic information for printed works, can polish a header for use in a catalog record, providing quality control and adding value to the header via version control. Publishers could even be involved in header production. Catalog librarians can create access points to the text by using the header. They would add authority control, as they do now for printed works, and they could possibly be enlisted for header quality control on some occasions. The system designers would design mapping mechanisms from headers to MARC fields for preliminary catalog records, and write programs so that users could browse the header effectively and selectively retrieve sections of a document.

The group's reporter summarized important points that the group identified for pursuing: Mapping the TEI header to MARC is an important aspect of their model. A practical handbook or guidebook should be created to explain standards for creating a TEI header, or header for any other electronic full text as well. A system must be set in place for file authentication. The numbering and referencing of different versions of texts should be standardized. Browsers which use the header for navigation must be developed to support different views of the document or to select a part of the document to view. The header should contain explicit information describing the non-SGML files included in the document. Finally, with more structure, the revision history section of the header could better contribute information regarding version control to the MARC record.

4.2 Group 2

Moderators: Peggy Seiden and Brian Schottlaender
Reporter: Brian Schottlaender

Participants

Randall Barry	Kate Harrie	Bob Scott
Michael Carroll	Annelies Hoogcarspel	Michael Sperberg-McQueen
Cynthia Gozzi	Mark Olsen	Stanley Yates
Peter Granda	Ann Sandberg -Fox	

This group discussed many of the same issues as Group 1, although their focus was slightly different. They began by discussing their expected outcomes and results:

- To define the role to be played by both small and large data providers—from those who encode only a single document to those who maintain databases of encoded documents, many of which are not SGML-conformant;
- To identify the role MARC records play as pointers to existing data services;
- To affirm or disaffirm the roles currently held by MARC, as a carrier for bibliographic data, and by SGML, as an encoding standard;
- To determine the relationship between MARC records and SGML encoded texts in how much they should coordinate and to what extent they should build on each other.

Most of this group's discussion time was spent debating the features of MARC and SGML. The group determined two options for their future coordination. The first, very much like that envisioned by Group 1, was to leave MARC and SGML separate, linked by bi-directional pointers. The second option was to combine MARC and SGML. The latter was envisioned in two ways. The first would be to "MARCize" SGML by expanding the TEI header file description section. This would mean using MARC conventions as the way that information is recorded in the header. The second would be "SGMLizing" MARC, or creating a MARC DTD, either to stand alone or to work in the TEI header. This would be an SGML tag set which encompasses the intellectual content of a MARC format, without regard to its syntax. For example, in TEI, only the tags <author> </author> are defined. In MARC, there are personal, corporate and conference authors, which would each be distinguished in a MARC DTD.

One advantage of combining MARC and SGML, regardless of which of these two possibilities was chosen, would be that any database of electronic texts would use only a single type of record structure. Also, by eliminating the need for bi-directional pointers, the likelihood for information loss would be reduced. However, one issue was raised which is important to the concept of combining MARC and SGML: It is not clear that there would be an unambiguous, one-to-one relationship between MARC and SGML. Regardless of whether the MARC record and the TEI

header should be left separate or combined, the group identified two issues to be resolved: (1) translation or mapping mechanisms, and (2) resources. The group concluded that the formats of the data are less important than the data that they carry or describe. Therefore, they felt it was desirable that SGML data elements be developed which are equivalent to those found in MARC, whether in a separate MARC DTD or in an expanded TEI header file. These various elements would be supplied both by the producers of the data, and by catalogers and publishers.

In response to one question asked of the discussion groups, Group 2 identified a number of pieces of information which would be needed to *describe, access, locate,* and *retrieve* remote electronic resources. The group also identified two more aspects of the question: what would be needed to *evaluate* and *use* those resources? Information needed for location and retrieval include host name, electronic address, path, and filename. The standard numbering systems described by Clifford Lynch earlier—URLs, URNs, and URCs—would be vital to location of an item. For access, compression information, login procedures, and passwords are needed. For some items, a contact person might be needed. In certain cases an operating system might be important. For identification purposes, other pieces of information would be important, some of which are already required in the TEI header, including title, publication and source statements. Subject analysis and classification are important for retrieval when a specific item is not known. Once it is retrieved, in order to evaluate and use the item it is important to know how it was encoded, what software is needed, and if an electronic document is "queriable" and what its querying protocols are. Identification of the creators is very important for authentication purposes, as are detailed version/release/edition information. Finally, the "physical" description of the item, either its size (such as 1 gigabyte of text) or its inclusiveness (such as an encyclopedia with photos) would be important. Specific forms of materials, such as images or sound, would probably require little more information beyond that listed here, other than the specific characteristics related to its format, so as to differentiate between an electronic map and an electronic photo, for example.

Like Group 1, Group 2 also spent some time trying to pinpoint the version/release/edition issues raised by electronic texts. The group agreed that different producers of information, such as libraries, data archives and data creators, define the terms differently, and that probably all of these terms cannot be equated; nonetheless it is important to be aware of the varying definitions. The group felt that providing a date/time stamp to an electronic document, on a broad scale, from year to nanosecond, would help identify different versions. Also suggested was a "hierarchy of infinite depth,"[8] where a version could be defined as "Release 1234.56789."

Finally, the group discussed the special needs of an electronic text that was only available in electronic form. The current TEI guidelines, based on previously printed materials, require a source description. The group felt that the source description requirement should be redefined and reconsidered in the case of new electronic materials, because it merely invokes a standard phrase with limited value in the source description section.

[8] Terminology is Michael Sperberg-McQueen's.

4.3 Group 3

Moderator: John Byrum
Reporter: Doug Kincade

Participants

Ed Gaynor	Hans Joergen Marker
Lisa Horowitz	Lynn Marko
Laurel Jizba	Keizo Oyama
Anita Lowry	Frank Sadowski
Clifford Lynch	Allen Renear

The conclusions of Group 3 were often very similar to those of the previous two groups. This group began by evaluating the relationship between MARC and TEI. They decided, like Group 1, that MARC and TEI must remain separate because their purposes are so different. The flexibility of the TEI header was viewed as a good thing in a continually revised document, because the header can then also be continually updated. However, this is exactly what prevents the header from being a viable access point, because it could not be a stable index. The TEI header is a good way to describe provenance, source, revision history, and full bibliographic information about a text. The TEI guidelines defining the header are heuristic, meaning that they are designed to prompt useful bibliographic information from an author who may not otherwise have considered providing it (or who did not have any other incentive to provide it). MARC, on the other hand, is granular and parsable. It is especially valuable for allowing access and describing the location of a document.

The group decided that a coexistence model was best. As discussed in Group 1, MARC should be the primary access point, with the TEI header as an intermediary step between access and the document. The group felt that a header separate from the document would be useful for browsing. The TEI header can inform the MARC record, possibly via direct mapping. MARC should be used for primary searching and authority records. The group also suggested that hits from the MARC record could be linked to the TEI header browser.

Envisioning the contents of the TEI header as a new source of information led to some other possibilities of its use with MARC. The group saw that new access points might be created from information in the TEI header. In particular, the group discussed “attribute access.” For example, the TEI header might include genre of the text or gender of an author, information that might be of interest to someone searching for specific electronic texts. This information is not accessible according to current cataloging rules.

Discussions of these concepts led to the question of what channels would lead from TEI to MARC and vice versa. The group’s reporter designed a conceptual map to show how different roles might be conceived leading from the TEI universe to MARC and back again. The

conceptual map placed the TEI header on the left, with users and file creators (comprised of authors, scholars, students, information technology specialists, and publishers) responsible for its creation and production. This side of the map was seen to represent the document's inner structure. The right side of the conceptual map contained MARC and access to information, representing exterior connections to the document. Librarians, especially catalogers, and some information technology specialists were listed as responsible for this kind of information. The focus of the map was in the center. Structures considered to mediate between MARC and TEI were listed here. Unique identifiers, such as URNs, were mentioned in this section, as they act both to validate version/edition information, useful to those in the left grouping, and to provide an access point for searching, useful to those in the right grouping. Attribute access, as described above, also crossed this middle ground. Software which maps between the two was listed in the center, as were standards for mapping and browsing.

The roles defined by this cross-over area were also discussed. Obviously naming authorities would be a bridge between the MARC and TEI worlds. Publishers too might lie in the center, because publishers have always been responsible for front matter, which contains much of the same information as the MARC record and the TEI header. The job of publishers has always been to create usable formats from scholars' information for the library community. Scholarly technology groups might also play a role by providing incentives to contributing scholars to include a well formatted TEI header in their documents. Finally, programmers would also be in this middle area, designing systems and software which would map from TEI to MARC, and creating browsers in the TEI header for access beyond the MARC record.

The group concluded by pointing out the importance of face-to-face meetings. Cyberspace can sometimes limit each individual to consider only to his or her local concerns, while face-to-face discussions with users of different perspectives broadens conceptual horizons. Alliances and lobbying within and across organizations are vital to the future of this kind of work as well.

4.4 Group 4

Moderator: Randall Jones
Reporter: Alan Morrison

Participants

Willy Cromwell	Daniel Pitti
John Duke	Laine Ruus
Dominic Dunlop	Joe Shubitowski
Rich Giordano	Don Thornbury
Peter Hirtle	Stu Weibel
Gregory Murphy	

Group 4 focused on two main issues: the relationship between the TEI and MARC, and proposals for future use. They determined that the information needed to describe, locate, access and retrieve electronic texts in the humanities is actually well defined within traditional cataloging methods and the TEI header guidelines. Beyond these, the group felt that it was impossible to determine what might be desirable without further experience using the guidelines. The idea from Allen Renear's talk, that people do not know what they want because of the restrictions of the infrastructure (in this case, the limited availability of standardly marked-up electronic texts), recurred in Group 4's discussions. The group felt that only by experiences and empirical work in the field using the header and working with MARC can we discover the shortcomings of the two systems. However, the group definitely felt that better mapping techniques from one to the other format was needed, requiring more structure in the fields of the header.

The argument that we do not yet know enough about work in the field to suggest improvements to TEI and/or MARC provoked much of the discussion about the future of TEI and MARC. The most important item in this future, the group agreed, must be education. Specific suggestions included increasing SGML awareness, teaching users how to encode automatically and how to use templates, and sponsoring hands-on workshops designed to show catalogers how to interpret the header. These kinds of educational forums would increase cooperation and navigation between the different fields.

The group also had some in-depth discussion of the nature of the problems involved in the relationship between MARC and the TEI. Again the lack of education on SGML was cited as a problem. Additionally, because of the extent of the TEI guidelines (TEI P3 is about 1300 pages), the group suggested simpler instructions on how to build software tools to work with the TEI guidelines and indicated a definite need for a "Reader's Digest" version of TEI P3. Software tools were desired which could standardize the wealth of information available over the Internet. Along these lines, software tools were also recommended to allow mechanical mapping of the TEI header to MARC.

Group 4 considered a few other points as well. One important issue revolves around the intellectual property rights of texts. Having a well marked-up text is useless if no one can see it or use it, as might be the case with many of the texts under copyright restraints in the British National Corpus. Second, the group stressed the importance of keeping the text independent from the software for the sake of long-term preservation of texts. The group emphasized the value of cooperation among different disciplines and organizations in order to allow for standards. They concluded with an emphasis again on the need for education and better tools and structure to work with the TEI and MARC together.

5. Conclusions and Recommendations

At the end of the workshop, all the participants gathered to listen to the reports from each break-out discussion group. After these were completed, a general discussion ensued. Many of the librarians in particular expressed much enthusiasm for a header on electronic files in general. TEI was recommended as a good model approach to a header on all electronic documents, and it was suggested that the TEI header be advertised to catalogers as the chief source of information for electronic texts. It was noted that currently many electronic texts are available that contain no explanation of the encoding they contain, or no citation from where the text was transcribed. As individuals add their own or competing analyses of a text, it will become increasingly critical to explain how the texts are encoded for other people to understand and use the texts. The header was praised for promoting the idea that when a text is created, it should be documented. The hope was expressed that scholars will consider the use of headers as a central part of their work. All of the different communities involved in the workshop must "sell" the idea of a header to be used on all scholarly work. In support of this, the publishing community could consider using the appropriate chapters from TEI P3 for encoding any scholarly work.

The importance of SGML, the TEI, and the header in cataloging was also declared. Training of librarians can expand the dialogue between some of the disassociated groups, and can extend the use of standards through library support. Training in SGML has already begun in some cataloging communities. It was commented, although not fully supported, that more automation of systems and even of cataloging could be possible with electronic texts. Also mentioned was the fact that TEI is only a small part of the increasing SGML standardization in the publishing industry for all kinds of texts, an additional reason for everyone to become conversant in SGML.

Some issues were raised regarding the discussion's focus on scholars. Although the group of users of electronic full texts in the humanities might be scholarly and specialized now, as electronic texts become more available to the general public, the needs of the general public should be considered in documenting. Within the scholarly community, meanwhile, production of electronic materials does not gain the respect that traditional print publishing has, a fact that is reflected in issues of tenure. The idea that a well marked-up text is a scholarly endeavor must be promoted. It was noted that preparation of primary material, as well as the writing of textbooks, are currently low prestige projects within academe, so emphasis should be placed on the intellectual ambitions of electronic texts. Additionally, we are just learning how to evaluate electronic texts, and we need time to learn to evaluate their computational properties as well as their intellectual content.

Susan Hockey, director of CETH, concluded by commenting that the workshop had raised awareness of the important issues involved in documenting electronic primary source materials in the humanities, and that participants should spread this awareness at their institutions. Other methods to spread awareness of the issues were discussed as well. Following the write-up of the technical report, it was strongly recommended that spin-off articles be generated in journals of all

the communities represented at the workshop in order to unite the different groups in the common cause of quality documentation of electronic texts. It was recommended also that connections be made with higher level administrators such as library administrators, academic deans, and publishing editorial boards to promote gatherings such as this workshop, and to support the work that will be required when some of these issues are resolved.

Appendix A

Program

Monday, May 16

3:00 – 6:00 p.m.	*Arrival*
6:00 – 7:00 p.m.	*Reception*
7:00 – 9:00 p.m.	*Dinner* Welcome Marianne Gaunt, Rutgers University
9:00 – 10:00 p.m.	Needs of Humanities Scholars Allen Renear, Brown University

Tuesday, May 17

7:30 – 8:45 a.m.	*Breakfast*
9:00 – 9:10 a.m.	Opening Susan Hockey, CETH
9:10 – 9:55 a.m.	SGML Michael Sperberg-McQueen, TEI/University of Chicago
9:55 – 10:40 a.m.	TEI Header Rich Giordano, University of Manchester
10:40 – 11:00 a.m.	*Break*
11:00 a.m.	Introduction Peter Graham, Rutgers University Libraries
11:00 – 11:45 a.m.	Networked Electronic Resources Clifford Lynch, University of California
11:45 – 12:30 p.m.	MARC Randall Barry, Library of Congress
12:30 – 2:00 p.m.	*Lunch*
2:00 p.m.	Introduction Annelies Hoogcarspel, CETH
2:00 – 2:20 p.m.	British National Corpus Dominic Dunlop
2:20 – 2:40 p.m.	Berkeley Finding Aids Project Daniel Pitti, University of California
2:40 – 3:00 p.m.	TEI Header and Cataloging John Price-Wilkin & Edward Gaynor, University of Virginia

Tuesday, May 17 (continued)

3:00 – 3:30 p.m.	*Break*
3:30 – 5:30 p.m.	Discussion in four break-out groups
7:00 – 9:00 p.m.	*Dinner*

Wednesday, May 18

7:30 – 8:45 a.m.	*Breakfast*
9:00 – 11:00 a.m.	Discussion in four break-out groups (continued)
11:00 – 11:30 a.m	*Break*
11:30 – 1:00 p.m.	Reports and joint discussion
1:00 – 2:00 p.m.	*Lunch*
2:00 p.m.	*End*

Appendix B

List of Participants

Randall Barry
Senior MARC Standards Specialist
Network Development and MARC Standards Office
Library of Congress

Judith M. Brugger
Catalog Management and Authorities Librarian
Cornell University

Arianne Burnette
Software Editor
Oxford University Press

John D. Byrum, Jr.
Regional and Cooperative Cataloging Division
Library of Congress

Michael Carroll
Senior Programmer/Analyst
Research Libraries Group

David R. Chesnutt
Professor of History
Univ. of South Carolina

Willy Cromwell
Principal Cataloger
Stanford University Libraries

Steven J. DeRose
Senior Systems Architect
Electronic Book Technologies, Inc.

John Duke
Assistant Director, Network and Technical Services
Virginia Commonwealth University

Dominic Dunlop
Computer Consultant
Luxembourg

Marianne Gaunt
Associate University Librarian for Research and Undergraduate Services
Rutgers University Libraries

Edward Gaynor
Head, Original Cataloging
Alderman Library
University of Virginia

Richard Giordano
Lecturer, Computer Science Department
University of Manchester

Cynthia I. Gozzi
Director of Technical Services
Stanford University Libraries

Peter Graham
Associate University Librarian for Technical and Automated Services
Rutgers University Libraries

Peter Granda
Research Associate
ICPSR
University of Michigan

Catherine I. Harrie
Senior Editor
Rutgers University Press

Peter B. Hirtle
Archives Specialist
National Archives and Records Administration
National Archives at College Park

Susan Hockey
Director
Center for Electronic Texts in the Humanities
Rutgers and Princeton Universities

Annelies Hoogcarspel
Catalog Librarian
Center for Electronic Texts in the Humanities
Rutgers and Princeton Universities

Lisa Horowitz
Project Librarian
Center for Electronic Texts in the Humanities
Rutgers and Princeton Universities

Laurel Jizba
Principal Cataloger
Michigan State University Libraries

Nancy R. John
Assistant University Librarian
University of Illinois at Chicago

Randall Jones
Dean, College of Humanities
Brigham Young University

Doug Kincade
Publisher, Electronic Products
Princeton University Press

Anita Lowry
Head, Information Arcade
University of Iowa Libraries

Clifford Lynch
Director, Library Automation
Office of the President
University of California

Carol A. Mandel
Director, Butler Library
Columbia University

Hans Joergen Marker
Senior Researcher
Danish Data Archive

Lynn F. Marko
Head, Monograph Cataloging Division
The University of Michigan Library

Alan Morrison
Oxford Text Archive
Oxford University Computing Service

Gregory Murphy
Text Systems Manager
Center for Electronic Texts in the Humanities
Rutgers and Princeton Universities

Mark Olsen
Assistant Director, ARTFL
University of Chicago

Keizo Oyama
Associate Professor, R&D Department
National Center for Science Information Systems, Japan

Daniel V. Pitti
Librarian for Advanced Technologies Project
University of California

John Price-Wilkin
Systems Librarian for Information Services
Alderman Library
University of Virginia

Allen Renear
Director, Scholarly Technology Group
Computing & Information Services
Brown University

Laine G.M. Ruus
Data Library Service
University of Toronto Library

Frank E. Sadowski, Jr.
Head, Catalog Department
University of Rochester Library

Ann Sandberg-Fox
Cataloging Consultant and Trainer
Vermont

Brian Schottlaender
Assistant University Librarian for Technical Services
University of California at Los Angeles

Robert H. Scott
Reference Librarian
Butler Library
Columbia University

Peggy Seiden
Head, Public Services
Skidmore College

Joe Shubitowski
Systems Project Manager
Documentation Projects
Getty Art History Information Program

C.M. Sperberg-McQueen
Senior Research Programmer
Computer Center
University of Illinois at Chicago

Sarah E. Thomas
Director for Cataloging
Collection Services
Library of Congress

Donald R. Thornbury
Head, Cataloging Division
Princeton University Libraries

Patricia Vanderberg
Librarian
University of California at Berkeley

Stuart Weibel
Senior Research Specialist
Office of Research
OCLC

Stanley Yates
Information Systems Specialist
Computing and Information Technology
Princeton University

Appendix C

Questions for discussion in break-out groups

1. What information is needed to describe/locate/access/retrieve electronic representations of primary source materials in the humanities from the perspective of

   (1) the user (research or teaching)
   (2) a) the catalog librarian
       b) the reference librarian
   (3) the compiler of the material
   (4) the publisher/distributor of the material
   (5) the implementor of a computer program which will process the material

2. Are there particular needs for specific forms of material, e.g. images, sound, text, etc.?

3. Is this information also appropriate for new material, that is material which is only in electronic form? If not, why not? What are the differences?

4. How can we deal with version or edition issues? What constitutes a new version and how is it documented?

5. How adequate are existing mechanisms (MARC, SGML/TEI or other)? To what extent do MARC and SGML have something to say to each other? How are they complementary? Do they need to be combined?

6. What are the systems implications? For example, how would you move from a catalog record to the information itself? What would such a system look like in ten years' time? How would we migrate to such a system?

7. What are the implications for librarians?

8. What are the implications for the user, the scholar?

9. What are the implications for the compiler of the material?

10. What are the implications for the distributor/publisher of the material?

11. What do you see as the biggest problem in this area?

Appendix D

For further information, see also:

Berners-Lee, T., et al. (August 13, 1994). "Uniform Resource Locators (URL)." Available at <URL:ftp://ds.internic.net/internet-drafts/draft-ietf-uri-url-06.txt>.

Gorman, Michael and Paul W. Winkler, Eds. (1988). *Anglo-American Cataloging Rules,* 2nd ed, 1988 revision. Chicago: American Library Association.

Hockey, Susan. (1993). "The ACH-ACL-ALLC Text Encoding Initiative: An Overview." Available by sending a "get TEIJ6.Memo" command to listserv@uicvm.uic.edu.

Hoogcarspel, Annelies. (April 1994). *CETH Technical Report 1: Guidelines for Cataloging Monographic Electronic Texts at the Center for Electronic Texts in the Humanities.* New Brunswick, NJ: CETH.

Hoogcarspel, Annelies. (1993). "Cataloging Electronic Texts and the TEI Guidelines," *Cataloging and Classification Quarterly, 16*(1), 126–127.

Hoogcarspel, Annelies. (1994.) "The Rutgers Inventory of Machine-Readable Texts in the Humanities: Cataloging and Access," *Information Technology and Libraries, 13*(1), 27–34.

ISO (International Organization for Standardization). (1986). *ISO 8879-1986 (E). Information Processing — Standard Generalized Mark-up Language (SGML),* 1st ed. 1986-10-15. Geneva: ISO.

ISO (International Organization for Standardization). (1988). *ISO 8879-1986/A1: 1988 (E). Information Processing — Text and Office Systems — Standard Generalized Mark-up Language (SGML),* 1st ed. Amendment 1. 1986-10-15. Geneva: ISO.

Mealling, Michael. (July 8, 1994). "Encoding and Use of Uniform Resource Characteristics." Available at <URL:ftp://ds.internic.net/internet-drafts/draft-ietf-uri-urc-spec-00.txt>.

Sollins, K. & L. Masinter. (March 26, 1994). "Requirements of Uniform Resource Names." Available at <URL:ftp://cnri.reston.va.us/internet-drafts/draft-sollins-urn.00.txt>.

Sperberg-McQueen, Michael & Lou Burnard, eds. (1994). *Guidelines for Electronic Text Encoding and Interchange.* Chicago: ACH-ACL-ALLC, Text Encoding Initiative. For information on obtaining these documents, send an "index TEI-L" command to listserv@uicvm.uic.edu.

van Herwignen, E. (1994). *Practical SGML,* 2nd ed. Boston: Kluwer Academic.

For further discussion, see the following lists on the Internet:

ETEXTCTR@lists.princeton.edu
Emedia@vax1.elon.edu
AUTOCAT@ubvm.cc.buffalo.edu
TEI-L@uicvm.uic.edu