The Librarian's Role in Building the Virtual Library

Taylor Fitchett, College of Law Library, University of Cincinnati, Ohio, USA

I. Introduction

We are entering a virtual age. The signs have been visible for over a decade. When my children were growing up I watched with concern as they preferred to interact with soccer balls on the computer screen rather than with neighborhood kids in our backyard. I imagined that a reduction in interaction with human playmates would create adult misfits in the *real world.* Only now, when they are adults, do I understand that the time they spent with joystick in hand was preparing them for their virtual future. My children and the wave of people behind them are ready for virtuality before it is ready for them.

Certainly a large number of people would be both technologically and psychologically ready for the virtual library if it were here today. Some define the virtual library as the library with no walls, no books, and no librarians. Others see the virtual library as a nexus for information activities, including e-mail, teleconferencing, newsgroups, listservs, etc. I have one of the more expansive visions of this library, one that encompasses access, delivery, and preservation systems, as well as education and communication systems. I envision one global library and one worldwide system of communication. In my mind the virtual library will have all the critical works of the libraries and research institutions in the world. This library will be accessible to anyone, anywhere, and at anytime. The vision includes all the communications and multimedia pieces, in real time, of course. The Internet is an essential component because it links resources and facilitates communications.

Librarians have been a part of the steady progression toward the virtual library, and we are also essential components in its development. In the years prior to the Machine-Readable Cataloging (MARC) era that began in the mid-1960s, libraries were organizing bibliographic information in a fashion that would lead to the international implementation of the MARC record, a development that revolutionized libraries. While the virtual library does not yet exist in its entirety, parts of it are in place, and its advent brings with it a fundamental change in the way libraries do business. Today, librarians are not only producing information for the Internet, we are trying to find ways to better index its content. Despite the problems in coping with the information revolution, we are excited by the promise of electronic information. While we share many of the same concerns over digital data as publishers and others in the information business, we also find ourselves involved with issues that are not foremost on the agendas of others, such as the authentication of documents, document citation and migration, the organization of huge bodies of data, information security, and user privacy. The preservation of our cultural heritage during the age of electronic information depends on coordinated strategies among libraries.

With no master blueprint on how to build a virtual library, libraries have begun to move collections into cyberspace. Special collections of photographs, manuscripts, and older literary works for which copyright is not a barrier, are among the first items from libraries to enter

cyberlife. There are a growing number of sophisticated electronic text projects in the States. Most notable for their standards in developing electronic text are the Model Editions Partnership, supported by Rutgers University and the Universities of South Carolina and Illinois at Chicago, and the Women Writers Project at Brown University. The Legal Information Institute at Cornell University is one of the few academic publishers of legal materials in electronic format within the States and an important colleague of the University of Cincinnati.[1] The goals and strategies of such groups have set the standard for the Center for Electronic Text in the Law at the University of Cincinnati, whose work, especially its work on the DIANA database of international human rights documents, is the focus of this paper.

II. The Center for Electronic Text in the Law

Four years ago the University of Cincinnati College of Law Library established the Center for Electronic Text in the Law (CETL), because we realized that the production of electronic text was an increasingly important part of our law school's operation. We also realized that we needed to look at information in new ways and develop new information paradigms to prepare to meet the predicted needs of the user. There was concern that although we increasingly relied upon the electronic medium for access to legal materials we had little input to either the development or use of that medium. With the support of the Dean of the College of Law and a few thousand dollars, in 1993 CETL was created.[2] It is fully integrated into our library's operations, and although it has several employees assigned on a full-time basis, all members of the library staff have responsibilities related to its work. CETL's mission is to promote legal scholarship through the understanding, management, and production of enhanced electronic text. The work of CETL can be divided into three categories:

1. Research on Electronic Text:

Research on methods of processing and managing electronic text is an ongoing objective at CETL. In the earliest stages of exploring hardware and software platforms[3] and various

---

[1]The Brown University Women Writers Project can be browsed at the following URL: "http://www.stg.brown.edu/projects/wwp/wwp_home.html." The Model Editions Partnership can be browsed at: "http://www.dlib.org/blib/november95/11chestnut.html." The Legal Information Institute's activities and future plans can be browsed at: "http://www.law.cornell.edu/papers/liirptf.htm"

[2]Nick Finke is the Executive Director of the Center for Electronic Text in the Law at the University of Cincinnati College of Law Library. Greg MacGowan is the Associate Director. Other staff members include text editors who are graduate students.

[3]CETL's decision was to build its databases on a UNIX foundation to take advantage of multi-tasking capabilities.

standards for electronic text, research was CETL's primary function. During this period we built a small prototype database for experimentation and demonstration, but we were not anxious to rush into text production without a review of the technology available at the time and an understanding of the standards associated with electronic data. Although CETL is more production oriented today, exploring new methods of electronic text management is still an important part of its agenda.

At the moment, we are looking at standards of encoding electronic text, most specifically using SGML (Standard Generalized Markup Language). We are examining both the costs and benefits of this labor-intensive markup system. SGML became interesting to us because it preserves text structure in an application-independent manner. The type of SGML that we are using is TEI (Text Encoding Initiative). TEI is an encoding standard for the humanities and social sciences that we have adapted for legal materials.

In addition to its research function, CETL acts as a text consultant for a number of universities who are developing technological approaches to publication. Consequently, an effort is made to stay current with new products and standards that relate to text management.

2. Publication Services to faculty and students

CETL also serves as a publication center for law faculty and students. We assist the law school community in building databases for research and teaching. Most of the assistance has been course related, and we have drawn on the resources of the University's Center for Academic Technology for consultation in instructional design. Faculty members, accustomed to the traditional method of teaching law, need guidance in preparing course materials for the online environment. Whether assisting in building a FolioViews database of case law or an HTML database for a class assignment, it is our goal to encourage the teacher to maximize the use of the instructional tool.

3. Electronic Publishing

CETL publishes two databases, both of which support centers of study that have been established at the University of Cincinnati College of Law. The Securities Lawyers' Deskbook supports the work of the Center for Corporate Law. The documents in this database are either obtained in electronic format from an information vendor, or scanned locally, and marked up in HTML. The other, DIANA, a collection of human rights documents, is an SGML database that supports the work of the Urban Morgan Institute for Human Rights.[4] DIANA is the more complex and interesting of the two databases, and its construction process merits description.

Before describing the techniques of building DIANA, a brief history of its origin is in order. The University of Cincinnati Law Library has a typical collection of legal materials for a law school. One collection distinguishes it from similar schools, its collection of human rights materials.

---

[4]DIANA is named in honor of Diana Vincent-Daviss, the late library director and human rights bibliographer from Yale Law School.

Although rather poorly indexed, the collection is, nevertheless, heavily borrowed from by other academic institutions. Several years ago a decision was made to create a database of human rights materials that would be accessible to researchers over the Internet. Many documents necessary to research in the field of human rights are difficult to obtain, especially in areas of the world where they are most needed.

At the same time that we were beginning our human rights initiative, we learned that a number of other institutions had a similar idea, and we decided to partner with these groups. We had several coordinating meetings and set up an advisory board of human rights scholars and activists. The members of the Board, who oversee the development of DIANA, have agreed that it will be a non-fee based library of human rights materials on the Internet. It was decided that the database will first contain the core instruments central to research in the field of human rights and expand to include difficult to obtain U. N. documents, briefs of the various human rights organizations, non-governmental organization information, and current awareness materials. As an international database DIANA will be in multiple languages, and it will ultimately combine the efforts of many organizations around the world.[5]

Unfortunately, at this time there has been no consensus among the contributors to DIANA concerning the technical standards for building the database. CETL, for example, stands firmly behind a set of principles for editorial markup that are not used by other contributors. It is our concern that the database be built to meet the scholarly editorial practices of today, that it be designed to limit the amount of information loss that could be encountered in future migrations, and that it be maximally transportable.

III. How the DIANA Database is Built at the University of Cincinnati [6] (see CETL Process flow chart)

1. Determining Intellectual Content

The first step in the process of building DIANA is taken outside of CETL with the selection of a particular title for inclusion in the database. As the project has grown, many people in the field of human rights have become interested in DIANA's scope of coverage. We have drawn on the work of bibliographers who are not on our staff, and we have consulted with scholars working in

---

[5]The institutions currently developing the DIANA database are The Orville B. Schell Center for International Human Rights at Yale Law School, The University of Minnesota Human Rights Center, The University of Toronto Law Library, Harvard Law School Library, the Urban Morgan Institute for Human Rights, and The University of Cincinnati College of Law Library.

[6]Additional information on CETL, including its technical processes can be found at the following URL: http://www.law.uc.edu/CETL.

the field of human rights to ensure the proper development of the intellectual content of DIANA.[7]

2. Acquisition of Materials

The acquisitions librarian locates and acquires materials for inclusion in the databases.[8] What makes the acquisition of documents for our electronic work challenging is that we strive to obtain original sources or the closest thing we can get to an original source. This means that if we are working on current electronic documents from the United Nations, we want to acquire them from the United Nations optical drive. If we are scanning hard copy of the Organization of African Unity resolutions, we prefer to get them directly from the OAU and not reprinted from some other publication.

The acquisitions librarian also acquires copyright permission for an electronic reproduction of a source document when it is required. America has not had a major revision of its copyright law since 1976, and there is much uncertainty about how the 1976 Act applies to digital media. What people seem to agree on is that any form of storage is a reproduction. There are so many parties interested in the transfer of electronic information, including copyright holders, librarians, publishers, reproduction rights organizations, various user groups, Internet providers, etc., that we are proceeding with extreme caution with any material not in the public domain.

3. Administrative Control of Text

From the moment the acquired document enters CETL it is tracked. Electronic documents and their paper counterparts are assigned accession numbers. The number always resides with the electronic copy, and if there is a companion paper copy, it is placed in a folder and archived under the same number. In this way, the paper copy is always available to the text editors who may have to consult the document later in the conversion process.

4. Text Conversion Process

The text conversion process, i.e. the process of taking an original document and putting it in electronic format, can be quite simple or very complex, depending on the intent for making the document electronically accessible and the format of the source document. Concerns with such things as document stability, transportability, and preservation of a source document make the conversion process complex and, consequently, expensive. The DIANA database is viewed as a

---

[7]Jeanette Yackle, Head of Reference for International and Foreign Law at Harvard Law School, Ewa Brantley, human rights scholar and activist, and Mariano Morales-Lebron, Head of Reference at the University of Cincinnati College of Law, have had primary responsibility for Cincinnati's selection of materials within DIANA.

[8]Cynthia Aninao, acquisitions librarian at the University of Cincinnati College of Law Library, works with CETL materials between 5 and 10 hours per week.

being a long-term resident within the virtual library, and therefore, must be built using the highest standards available to us today for the management of electronic text. A document of transient importance can be converted simply and inexpensively in any number of ways, including having it scanned into PDF, marked up in HTML, or even left in a flat ASCII format.

The path that a document follows once it arrives in CETL differs depending on a number of variables. If it arrives already in an electronic format, as most of DIANA's United Nations documents do, a number of processing steps will be eliminated. If, however, the source document is in paper form a decision must be made as to whether it will be digitally imaged and OCR'd or whether it is best sent to a vendor who will rekey the text. Unless the original text is very high quality, we have found that it is considerably cheaper, roughly six times less expensive, to have the text rekeyed.

As the Center for Electronic Text in the Law has grown and learned more about processing text, we have reduced labor costs. One way to reduce the cost of processing electronic text is to find additional ways to automate the process. We continue to explore new software and we write small programs to improve existing software. Another way of reducing labor costs is simply to do less text processing. At CETL, that translates into doing less markup. The level of markup needed for a particular text has caused some controversy among the text editors in CETL. Once you actually begin analyzing a particular text it is quite easy to separate it into distinct elements that could be marked up. But when we considered the time spent in building the document's metadata file and added to that the time spent in the most basic markup, to do such things as identify the document structure and set up the links to other documents, it was clear that we had to reduce the labor costs of the markup process.

5. Imaging of Source Documents

Once a document is acquired, if the original is a paper document, it is prepared for photocopying. Usually it is the photocopy as opposed to the original document that is scanned. A digital image, that will later be tied to ASCII text, is then created. CETL scans documents at 400 dpi for optimal creation of text.

6. Image to ASCII

Next the image is turned into text using Xerox Imaging Systems' ScanWorX. The raw ASCII text is edited and spell-checked. The text is marked with styles in Rich Text Format (RTF) using StyleBuilder.[9] Unless the text of the original is very clear, a great deal of effort is needed to correct the errors made in translating the image into ASCII text. Older text, heavily footnoted materials, or text with tables have proven difficult to scan. We have found that converting text with an optical character reader is usually too expensive, exceeding $12 per page, and use the method infrequently.

---

[9]RTF allows importing from and exporting to SGML compliant software.

When images do not scan well due to the poor quality of the original, the document is sent off-site to be double-keyed.[10] The conversion company will add basic SGML markup to the ASCII text after an appropriate DTD (Document Type Definition) has been generated. The DTD must be perfected to assure the best return of data.

7. Document Markup

Once the document is in electronic form, CETL turns its focus toward adding value to the electronic characters. An important means of adding value and preserving information is by marking the text in Standard Generalized Markup Language (SGML). Text is fed through a software program that does some basic document markup automatically. Fast Tag had been used for this purpose, but the new middleware of choice is Omnimark. This raw, tagged text must be enhanced by text editors in order to create proper SGML. Text is placed in an SGML database that configures the SGML tags to operate as database fields, allowing accurate search and retrieval. The text can be pulled out of the database into many SGML-compliant applications.

8. Standard Generalized Markup Language

Text markup is so critical to the management of electronic information that a small digression to explain its importance is merited.[11] In the future, we do not want to discover that we have limited people's ability to use the electronic text we have created. Software changes rapidly and we risk information loss if we do not protect it during the many migrations it will surely make.

The University of Cincinnati Law Library was among the first law libraries to use SGML as a distribution and preservation medium for legal information. SGML is an international standard (ISO 8879) adopted in 1986 for the description of marked-up electronic text. It is a markup language, i.e. a set of instructions for encoding text. Encoding text is a means of making explicit the interpretation of text, e.g. identifying the author or title of a publication.

SGML provides a standardized structure to assign attributes to a text and define the structure of the text. This type of tagging is critical for operations such as field searching and search limitation as well as page presentation and layout. SGML is descriptive rather than procedural. It can be used with many different applications, and it is used widely by publishers and by scholars who are building databases.

SGML allows the separation of text and structure. Because of this ability SGML documents can be interchanged among many systems in many ways. The need to use electronic texts in new applications helped drive the development of SGML. Interchanging documents with minimal

---

[10] The company, Input Center, 320 N. Michigan, Suite 404, Chicago, IL 60601, charges 85 cents per thousand characters (approximately $2 per page) which includes the markup.

[11] Robin Cover's SGML Web Page, a comprehensive site for information on SGML, has the following URL:
http://www.sil.org/sgml/sgml.html.

information loss as they move from application to application will preserve content. SGML is non-proprietary; it is vendor independent and application independent.

The type of SGML employed by CETL is called TEI (Text Encoding Initiative). *The TEI Guidelines for Text Encoding and Interchange* are a markup developed by scholars, librarians, and those interested in computing for use with literature in the humanities.[12] CETL is currently working with the Legal Information Institute (LII) at Cornell University to expand the TEI Guidelines for use with legal materials.

To create a final archival product, CETL uses TEI markup to encode the document's significant structural features, usually down to the paragraph level. Quotes, underlined words, tables, and other features that fall within paragraphs are also marked up. Markup includes identification of the basic reference unit so that it will be possible to create hypertext links to it later on. If the document exists in several languages, as often happens with United Nations materials, markup is added to indicate parallel points in the various language versions. When needed, pagination of the original paper document is indicated in the electronic version. Where there have been hard copy source documents, links are made to the digital images that were created earlier in the process. This completes the creation of the archival electronic text, and all subsequent distribution of this text is done from the SGML document.

As the virtual library grows, so, too, will the need to extract meaning from vast amounts of textual data. In building the DIANA database we are constructing a foundation for hundreds of millions of documents, not just for the relatively small number of documents now in the database. Based on our knowledge of the search and retrieval software existing today, SGML markup is an indispensable component of a research database.

9. Assigning the Metadata to the Document

There is a growing awareness among the information industry of the importance of assigning an appropriate amount of information to a particular document, or object, to define and index it during its electronic journey. One of the most critical roles that the librarians who work in CETL play is in the assignment of metadata, data describing data, to the document. CETL has automated much of this process, but the catalogers still have a bit of work in deciding how bibliographic data will be managed in the TEI header. CETL's catalogers are involved in building the structure of the DIANA database and in designing the electronic header that travels with all of the DIANA documents created by CETL.

The TEI header has four parts: 1) a file description, containing a full bibliographic description of the source document. Subject matter keywords chosen from standard thesauri can be indicated

---

[12] The TEI Guidelines for encoding and interchanging electronic text were developed under the direction of the Association of Computers and the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing. The latest version of the TEI Guidelines (TEI P3) was published in 1994.

here. In the case of DIANA we use HURIDOC[13] subject headings as well as MARC headings. 2) an encoding description, describing encoding practices 3) a text profile, where, for example, CETL explains how text ambiguities were handled and who did the work 4) a revision history, where any subsequent changes made to the text would be indicated.

10. Electronic Text Distribution

Once the SGML document is created it can be delivered in three ways. The first method is through down-translation to word-processing formats by means of SGML Hammer, a tool from Interleaf Avalanche. Hammer allows the user to create an RTF document from the SGML and save the RTF document in a word processing application.

The second method of delivering SGML requires the receiver to use an SGML viewer, such as the Panorama viewer. This viewer from SoftQuad may be employed by a Web browser when it encounters SGML. It downloads the text, the DTD, and the stylesheet to the client machine. This process may be cumbersome to many users, so other means of delivery must be available.

The final method of delivery is through DynaWeb, an HTTP server plug-in that allows SGML documents to be viewed on the Web in HTML. The user needs only Web access and no additional applications to view the document, but much of the SGML added value is lost.

In order to have access to the added value offered by SGML a new markup is being developed, Extensible Markup Language, XML, a simplified form of SGML. The goal of XML is to enable SGML to be served, received, and processed on the Web in the way that is now possible with HTML. When XML becomes available, CETL will use it to replace the current DynaWeb and Panorama delivery vehicles.

IV. Conclusion

In this brief overview of the work of the Center for Electronic Text in the Law, a glimpse of one of many initiatives coming from libraries around the world to build a virtual library has been offered. I have no realistic estimate of what it will cost to make this library a reality or how long it will take other than to say that it will cost trillions of dollars and take a long time to get the billions of documents of the world's research institutions alone converted to electronic text. America's National Archives houses 6 billion documents, and it would be safe to say that it will take decades to get the documents from this single institution into electronic form. Perhaps information specialists of the future will decide that such a retrospective endeavor is too ambitious.

---

[13] HURIDOCS is a network of non-governmental organizations and individuals concerned with human rights documentation who are striving to build one set of information standards, HURIDOC Standard Formats. Their guidelines are based on the *Anglo-American Cataloging Rules* and include features such as geographical terms and codes, human rights indexing terminology, and guidelines for recording the names of persons.

Those who have limited experience in the production of electronic text are shocked at the expense of building an electronic library of research quality materials. They are also surprised that the costs do not lie primarily in the acquisition of hardware, software, or even in the cost of information itself. The majority of dollars are spent on personnel to understand the technology, acquire the information, and then process it. Despite the cost of new technologies, library budgets are gradually being reallocated to accommodate them as the concept of access to information, versus ownership of information, is promoted by librarians. If budgetary constraints are used as justification for not exploring new ways of managing information, librarians may find that they are less relevant to the future.

While almost all of my colleagues in law libraries are involved with the Internet to locate information and create home pages, a much smaller number are actively engaged in building research quality databases for the Internet. Justification of funding for a text center, such as CETL, within a law library has been requested by law professors, law deans, and law librarians alike. Setting a new agenda in an organization requires months, or perhaps years, of groundwork with those who control the budget, the administration, and those who do the work, the librarians. Administrators, who are themselves under financial constraints, want to understand the cost-benefits of innovation. Text conversion projects are expensive and difficult to justify through conventional methods. Librarians who see the transition from hard copy to electronic delivery of information as an omen for the disintermediation between librarian and end-user may not eagerly facilitate the change. In most libraries it is the existing workforce who will implement change, so it is critical that they share a common vision for a new direction and possess the skills to achieve their goals.

While the librarian's part in building the virtual library is yet to be determined, early indicators are that the role will be significant. Librarians will not be the only players building the virtual library. Authors, publishers, computer scientists, and others will have a significant impact on its character. Librarians, who have several decades of experience in the creation and maintenance of online bibliographic databases, understand that the next logical progression is to link those existing databases to full-text. However, to date, they have given little serious attention to the residence of scholarly information on the Internet. There is enthusiasm among librarians about the Information Highway for many reasons, including the fact that information can be made accessible to remote users, that multiple users can access the same information simultaneously, and that powerful search tools offering full-text indexing can be found. At the same time there is the gnawing realization that information on the Net is not secure, well-preserved, well-organized, or necessarily authentic.

The journalist and writer, Bruce Sterling, offers one of the more pessimistic views of the digital revolution:

> Computers swallow whatever they can touch, and everything they swallow is forced to become as unstable as they are. With the soaring and brutal progress of Moore's Law, computer systems have become a series of ever-faster, ever more complex, and even more elaborate coffins.[14]

---

[14]Bruce Sterling, ***The Digital Revolution in Retrospect***, 40 no.2Communications of the ACM 79 (1997).

Certainly it is possible to envision us all as the victims of the information revolution, drowning in data today without the certainty of having access to any of it tomorrow. But it is equally possible to envision us as the beneficiaries of the same revolution, having learned to manage information through the use of technology. Computers can enhance learning, giving us new ways of looking at information. Sophisticated retrieval software will simplify our quest for information by helping us evaluate search results from vast data repositories. However, lack of foresight on the part of the library profession, which has been charged by our society with the management of information and the preservation of our cultural heritage, would certainly drive a nail into the information coffin. But a look at the twentieth century library system reveals a highly structured and successfully managed network of information centers. As these networked entities begin to meld into the virtual library, librarians will develop and apply the appropriate management standards to the electronic medium to ensure the viability of yet another mode of information transfer, whatever its ultimate duration may be.

CETL PROCESS