Evaluating Electronic Texts in the Humanities

SUSAN HOCKEY

ABSTRACT

The number of electronic texts in the humanities is growing fast and many libraries are seeking to acquire them from various sources or to provide access to them. Building, in part, on the experience of those scholars who work with electronic texts in literary and linguistic computing, this article surveys some issues which libraries may need to consider as they begin to establish collection-development policies for electronic texts. An overview of existing texts and applications is given, which leads to a discussion of markup schemes. The Text Encoding Initiative's proposals for documenting electronic texts are surveyed, and the article concludes with a discussion of software and access tools.

INTRODUCTION

Electronic texts have been used for scholarly research in the humanities for the past forty years or so ever since Roberto Busa began work on his Index Thomisticus in 1949. However, it is only in the last three to four years, and particularly with the advent of the Internet, that humanities electronic texts have moved into the center of the scholarly arena as libraries begin to collect them and provide access to them. In the humanities, as in other disciplines, electronic textual resources offer many more possibilities than print, but, in general, libraries do not yet have any well-established practices for collecting and handling electronic texts as they have with print material. Shreeves (1992) discusses some of these questions from the
perspective of the librarian, but there is a need also to look at what humanities scholars might want to do with the texts.

Electronic text is used here to mean primary source material in the humanities rather than journals and reference works. Such texts may be literary works (prose, verse, drama), historical papers, letters and memoranda, charters, papyri, inscriptions, and the like. The source material may be in any natural language and may be in print or manuscript form. The focus of this article is also on transcripts of text rather than digitized page or manuscript images. Images provide an exact reproduction of the original so that marginalia, annotations, parallel texts, illustrations, and the like are readily available. They can be used for access and preservation but the text cannot be searched or otherwise manipulated. A transcript of a text allows many more novel possibilities for research and teaching and exploits more fully the capabilities of electronic materials. In the future, a combination of image and text may well form the basis of the electronic library, where it will be possible to search the text and retrieve the image.

THE PRESENT SITUATION

The picture in the early 1990s is one of many humanities texts in many different places and in many different formats. The Georgetown Catalog of Projects in Electronic Texts lists over 300 institutions which hold electronic texts but not the texts themselves. From sources such as The Humanities Computing Yearbook 1989-90: A comprehensive guide to software and other resources (Lancashire, 1991), journals, and proceedings of annual conferences on humanities computing, the number of existing electronic texts in the humanities can be estimated at many thousands. The Internet gives access to a fraction of these, and the existence of most of the others is only known from articles which describe their use in specific projects.

Most of these texts are held by individuals or by research institutes (mainly in Europe) which have compiled them for their own research purposes. Examples include the Istituto di Linguistica Computationale in Pisa, and the Institut für Deutsche Sprache, Mannheim. For a variety of reasons, most of the collections of these institutes are not available for others to use. The few exceptions include many of the texts which were compiled for the Trésor de la Langue Française at Nancy, which are now available from ARTFL (American Research on the Treasury of the French Language) in Chicago. The texts compiled for the Responsa Project at Bar-Ilan University are now available on CD-ROM as the Global Jewish Database, and the collection of Early Christian Latin at Louvain-la-Neuve has now been published as the CETEDOC CD-ROM.
The Thesaurus Linguae Graecae (TLG) was the earliest systematic attempt to create electronic versions of the complete literature of one language (Ancient Greek) and its 60 million word task is now almost finished after twenty years of work. The Packard Humanities Institute (PHI) has completed a complementary collection of Classical Latin which is about 8 million words. Both of these are distributed on CD-ROM.

The largest general purpose collection of electronic texts is the Oxford Text Archive (OTA), which was established at Oxford University Computing Services in 1976 in order to prevent texts from becoming "lost" once their compilers had finished with them. The bulk of its collection comes from donations from individual scholars. It is committed to maintaining any text which is deposited in it but does not actively pursue material to be added or correct errors within the texts. It now has some 1,200 texts in about thirty languages and makes these available at nominal cost provided that the compiler has given the appropriate permissions. Little information is known about the source of some OTA texts, and the OTA takes no responsibility for the accuracy of the texts. Some texts are available by FTP from <black.ox.ac.uk>.

It is estimated that about 95 percent of existing texts are plain text files—that is, ASCII files which are not indexed for any specific software. Those who use them must acquire or develop suitable software programs, depending on the nature of their application. Various software programs for humanities electronic texts are in widespread use, notably the Oxford Concordance Program (OCP), Micro-OCP (a PC version); TACT and WordCruncher (interactive text retrieval programs) which all provide some basic facilities as well as more sophisticated tools tailored to the specific needs of the humanities.

The remaining 5 percent of texts are what can be called packaged products, where the text has been indexed for use with specific, often proprietary, software and cannot be used for any other purpose. Most of these products, at present, are on CD-ROM (e.g., the CETEDOC library of Christian Latin Texts, the WordCruncher Disc of American and English literature, the New Oxford English Dictionary on CD-ROM, and the Global Jewish Database). Libraries which provide these packaged resources will generally find that the support costs are not insignificant. Almost all use their own query language which in most cases is not intuitive. It takes some time to gain a good understanding of the full potential of many of them. They provide complex search facilities because the texts themselves are complex and scholars want to study them in many different ways. However, most of these products do have manuals which document the source
of the text as well as how to use the programs, which is more than can be said for some texts.

The Internet gives access to ARTFL and to the Dartmouth Dante Project (DDP), which includes the text of *The Divine Comedy* and major commentaries. ARTFL uses software developed by its own team, the second version of which is based on UNIX utilities and is not particularly easy to use for those not familiar with UNIX. The DDP uses BRS-Search with a user-friendly interface. It can perform flexible searches, but scholars who use it extensively will begin to see the limitations of applying a commercial text-retrieval system, which is document-oriented, to complex humanities texts, where it is not clear what constitutes a single document.

At present, it is rare for several different electronic versions of a work to exist. The Bible and Shakespeare are the exceptions. Comparisons of these can help in establishing design principles for better electronic texts. Bolton (1990) reviews three electronic versions of Shakespeare and tools to access them and gives a detailed evaluation from the perspective of a scholar in English studies. This essay highlights the relevance of complex tools for what are complex texts and the need to provide good documentation for them. Most of the current electronic versions of the Bible seem to be intended more for the popular market and only one or two would be really suitable for scholars and students in religious studies. As yet, there are few comparative reviews similar to Bolton's, but soon there will be more versions of electronic texts to choose from and more evaluations are sorely needed. Likely candidates might include a comparison of the texts of J. P. Migne's *Patrologia Latina* published by Chadwyck-Healey.

Given the present situation, how can a library evaluate electronic texts now? What makes a good electronic text that a research library would want to acquire or access? The popular market for electronic texts is growing fast. We are already witnessing different collections of electronic texts which are intended for popular, rather than scholarly, consumption. How can a research library ensure that a text is suitable for its collection? What do libraries need to know to make decisions on what to collect and how to provide access to their collections?

Some very basic questions which should be asked include: How are patrons going to use the text? Is it a text which requires software from elsewhere, or is it a complete package? If it requires software, what is the best program? What facilities does it provide? Are these facilities suitable for scholarly applications? What source text was used? What guarantee is there that the text is accurate? What markup scheme does it use? How will the electronic text be supported in
the library? What can the text do for patrons that print materials cannot do? Some understanding of what is involved in creating an electronic text and of basic techniques and applications in literary computing is helpful in order to begin to answer these questions.

**Creating Electronic Texts**

Most existing texts have been created by keyboarding in one way or another. Ones which were created many years ago, including the TLG as it is stored on the CD-ROM, are entirely in uppercase letters. Some of these have been converted to upper- and lowercase by software and thus often do not begin sentences with uppercase letters. Much keyboarding has been done by individuals who needed to create an electronic text for their own research purposes. They may have consciously or unconsciously edited the text to suit their own needs and understanding and have possibly not documented these changes. They may also genuinely have made mistakes which went unnoticed in proofreading. If the text has been keyboarded professionally, it may be less likely to contain mistakes, as in the case of the TLG where very few errors exist. However, if significant cost has been incurred in creating the text, it is perhaps less likely that the text is widely available for others to use. At present, the texts which are most widely available are often those that have been created by individuals. Experience has shown that accuracy must not be assumed.

Optical character reading (OCR) has been used to input a variety of humanities texts. The capability of OCR software has improved somewhat, but OCR is not yet able to handle early printed books, manuscripts, some types of newspapers, and other material printed on poor quality paper. A text which has been input by OCR will need thorough proofreading even if the initial scanning appears to be very good. Claims of accuracy rates of 99.9 percent in effect mean two or three errors per page, which is far more than one would expect to find in a printed book. Many existing texts which have been input via OCR have not been proofread well. Typical errors include confusion of e and c, h, and b and the number 1 and letter l, as well as words run together or spaces inserted in the middle of words. Extraneous matter on the page—such as blotches on a photocopy—will be read as apostrophes or commas. But, even if the letters have been recognized accurately, it is becoming increasingly clear that optical scanning yields only part of what is needed to create a useful electronic text. It gives a physical representation of the text, which can be ambiguous without additional information. For example, a word which is in italic could be a foreign word, part of a title, or an emphasized word. These distinctions need to be made for a retrieval program to be useful, but they can only be done by adding information
to a text after it has been scanned. When a text is keyboarded, this information can be embedded in the text at the time of capture in the form of encoding or markup tags.

As with print materials, the choice of source edition is important for the academic acceptability of a text. Many existing texts have been created from out-of-copyright editions because their compilers have not been able to obtain copyright permissions for newer editions or have not wanted to ask permission for fear of becoming embroiled in legal issues. It is often the case that more recent editions have greater scholarly value and would be more appropriate for research use. It is to be hoped that these copyright issues will be tackled and resolved in the future rather than being avoided, which seems to be the case at present. Another factor which has determined the choice of source edition is its suitability for scanning. Again, the text which can be read best on the scanner may not necessarily be the edition with the best scholarly value. Good intentions to edit the text so that it conforms to a better edition are sometimes not carried out. When shown electronic texts, scholars who are skeptical about their value often voice their concern by criticizing the choice of source edition, which is, after all, something they understand from traditional scholarship. The lack of good scholarly texts in electronic form has seriously hampered the development and acceptability of full-text applications in humanities research and teaching. This situation is not being helped by various projects which use the Internet to announce, and make freely available, texts which do not appear to have any particular scholarly value.

USES OF ELECTRONIC TEXTS IN THE HUMANITIES

Certain methodologies and techniques for literary and linguistic computing have been well understood for some time. Hockey (1980) gives an overview of applications, many of which are still current. Butler's (1992) collection of essays is also a useful source. The journals Literary and Linguistic Computing and Computers and the Humanities and the proceedings of various literary and linguistic and humanities computing conferences also give some background (Miall, 1990; Hockey & Ide, 1991).

Concordances and text retrieval have formed the major application areas in literary and linguistic computing. A concordance is an alphabetical list of words which shows all the instances of a particular form, allowing the scholar to examine them in fine detail. Text retrieval gives instant access to occurrences. The first and most obvious application of these is as a reference tool. In the humanities, questions such as, Does this word ever occur in this text? are as common as, Find a text about this topic. For the former, the text
must be absolutely accurate otherwise the user cannot be sure whether the word exists or not.

Concordances have been used as a basis for stylistic analyses and even for studies of disputed authorship. It has been shown that the style of an author, or even a genre, can be characterized by the use of function words, that is, words which authors share in common with their contemporaries. In their study of the *Federalist Papers*, Mosteller and Wallace (1964) showed that the use of words such as "whilst" or "while," "enough," and "upon" in the disputed papers followed that of Madison rather than Hamilton. Burrows (1987) used a concordance program and some simple statistics to show that the thirty most common words in the novels of Jane Austen can distinguish the "idiolects" of the different characters in her novels. Kenny's (1978) work on the *Aristotelian Ethics* is another classic example of traditional literary and linguistic computing techniques where a study of particles and other common words in Greek shows that the three books which appear in both the *Nicomachian Ethics* and *Eudemian Ethics* of Aristotle are more like the *Eudemian Ethics*. There are many other similar studies, all of which are based on common words which therefore need to be indexed.

Other computer-aided research has concentrated on the production of new critical editions in print form and now also in electronic form. Collation, concordance, and statistical tools can help the scholar establish the text and provide information for the commentary and other annotations (Robinson, in press). Other studies have included programs to analyze sound patterns and correlate these with the sense as, for example, in the *Divine Comedy* (Robey, 1987) and in Homer (Packard, 1974). Most kinds of research, which are based on lexical analysis, are suitable for computational techniques, provided that it is understood that the text is viewed as a sequence of graphic forms. Programs for automatic lemmatization (putting words under their dictionary headings), syntax, and morphological analysis are not yet widely available, and, in any case, those that do exist are never completely accurate and require manual verification of the results.

Hypertext applications have also become popular in the humanities (Delany & Landow, 1991). Images and sound can be linked to texts. More importantly, hypertext does not require the text and ancillary material to be constrained into a rigid structure such as a relational database. The data can be as flexible and extensible as needed, thus allowing the scholar to add more information or reorganize existing material as he or she learns more from working with an electronic version of it. The best known humanities hypertext is Perseus, which was developed by a consortium of institutions based
at Harvard (Mylonas, 1992). Perseus goes far beyond the text. It is a multimedia encyclopedia of Ancient Greek literature, archaeology, geography, history, and culture. Besides the works of major authors in Greek and English, it contains photographs of vases, sculptures, coins, buildings, and archaeological sites as well as an encyclopedia, historical overview, and Greek/English dictionary. Although Perseus is currently available on CD-ROM, the Perseus team sees the network as the future means of accessing the database and have designed it so that the individual components can easily be imported into other systems.

Retrieval and other applications on humanities texts can be complex simply because of the nature of the texts and the fine detail in which they are normally studied. A text may contain several different natural languages, some of which may be in different scripts or use different alphabets for sorting words. Examples include parallel texts of the Bible or Middle English texts which contain sections in Latin with citations of Greek or Hebrew words. Users of the texts must be able to identify which sections are in each language and to index them separately. Variant readings or spellings may be indexed. Quotations from other texts may need separate treatment. Punctuation is important in early printed books and may also need to be searchable. Studies of morphology in inflected languages or of rhyme can benefit from a reverse index where words are alphabetized according to their endings.

The canonical referencing scheme or logical structure of many humanities texts is complex, yet it needs to be represented in an electronic version. Depending on the type of literature, there are many different subdivisions of verse texts (stanzas, verses, books, quatrains, and so on). In simple terms, a play is divided into acts which are themselves divided into scenes and speeches. It also has a cast list and stage directions. However, a play may also have another referencing scheme which is based on pages within a printed edition. Line numbering may be in relation to the pages or sequential throughout the text. Printed editions of early manuscripts may also have two parallel referencing schemes—pages and lines in the print version as well as folios and lines in the original. All of these should be accessible to the scholar working on the text and therefore need to be identified or encoded within the text. An overview of some of these issues and the need for encoding to handle them is in Sperberg-McQueen (1991).

**MARKUP**

Markup or encoding makes explicit for computer processing those features of a text which are implicit for the reader. A text without
markup is like a bibliographic record which is not divided into fields. Markup is needed to identify the different elements of the referencing scheme as well as to distinguish among features which would otherwise be ambiguous and to encode features of interest. The period (full stop) is also used in abbreviations or as a decimal point in numbers. Some programs delimit concordance citations by orthographic sentences—i.e., by all the text up to a period—and so, without additional markup, abbreviations and decimal points would be erroneously considered as sentence boundaries. Quotations from another author or text need to be identified by their source. For studies such as the Burrows's work on Jane Austen cited earlier, markup is needed to separate the dialogue from the narrative in the text and to identify the speaker for each section of dialogue. In a play, markup could encode the change of speakers and stage directions as well as the logical structure. For further discussions on markup and scholarly text processing, see Coombs et al. (1987) and Renear et al. (1992).

A text without markup can only be used for very simple applications. One analogy is trying to perform functions such as sorting and searching on a bibliographic record which does not have field divisions. For textual analysis, this amounts to making a simple alphabetical list of words, counting the word frequencies, and performing very simple searches. None of these will be completely accurate for detailed analyses. A look at various versions of Shakespeare which are available over the Internet will immediately show the problems. Act and scene numbers are not marked up in any way and so will lead to word counts which include all the incidences of the word “Act” as the act number within those of “act” as a verb or noun used in the normal way. Roman numerals used as act and scene numbers are more problematic. Act I will be assumed to be an occurrence of the personal pronoun I. Even the WordCruncher CD-ROM suffers from this problem. The simple searches will retrieve one or more surrounding lines of context. With a prose text, the reader may want to reformat the lines as on a word processor when the margins are changed. In verse, the lineation is fixed and must not be reformatted. When a text is entirely in verse, one can allow for this, but texts which are mixed verse and prose need to have markup to show the difference. Words which are not in the main language of the text also need to be encoded so that they can be distinguished. Examples include English “vale” and Latin “vale” (farewell), English “pain” and French “pain” (bread).

Many different markup schemes have been developed for humanities electronic texts over the last forty years. Of these, the most notable are COCOA and its variants. COCOA was first devised for an Archive of Old Scots Texts in Edinburgh in the early 1960s.
(Aitken & Bratley, 1967) and is described fully in the Micro-OCP manual. It provides a way of encoding the canonical referencing structure of a text including parallel referencing schemes and can also be used for other features such as stage directions, editorial comment, and so on. It is used by most of the major text-analysis programs in current use in the humanities, notably the Oxford Concordance Program (OCP) and (in extended form) TACT. The Thesaurus Linguae Graecae developed its own markup scheme, called beta code, which has also been used by other projects in classics and religious studies. The retrieval program WordCruncher also has its own markup scheme. Many existing humanities electronic texts are encoded for use by these programs.

Typographic markup is also needed to print or display a text so that it is more easily readable. Even simple word processing programs include features such as italic, bold, and so on to highlight sections of a text to draw the reader's eye to them. A parallel set of markup schemes was thus developed for printing and formatting, most notably TeX, TROFF, and later various word processors, such as WordPerfect, where the markup is exposed by the Reveal Codes function.

The result of this plethora of markup schemes has been described as chaos (Burnard, 1988). By the mid-1980s, experience showed clearly that markup is essential for good quality texts, but no scheme had wide acceptance. Each scheme was designed for a specific project or application. Most schemes were poorly documented and had no provision for extension or were not otherwise sufficiently flexible. Much time was wasted on converting from one format to another. None of the existing markup schemes was suitable for adoption as a standard.

In 1986, the Standard Generalized Markup Language (SGML) became an international standard (van Herwijnen, 1990). SGML is not, in itself, an encoding scheme. It provides a syntactic framework within which descriptive information about an electronic text can be encoded. The principle of SGML is descriptive, not prescriptive—that is, it describes the structure of a text. It enables the word which is seen to be in italic to be described as part of a title, or a foreign word, or an emphasized word, or whatever the encoder wishes. At a very basic level, SGML views a text as being a collection of objects called elements. These may be chapters, pages, words, lines, stanzas, or whatever the user wishes. The set of elements for a particular text or group of texts and the relationship among them is defined in a document type definition (DTD). The DTD has a formal structure. It can be read by a computer program called an SGML parser which validates the markup in a text or by other SGML-based software.
which operates on the text. SGML provides a method of encoding which addresses many of the intellectual issues which previously used encoding schemes have not. A further advantage is that it also provides links to material which is not ASCII text—e.g., sound and images—which are likely to become increasingly important. Its one disadvantage is that it views a document as a single hierarchic structure and has no easy way of dealing with the multiple parallel referencing schemes which appear in many humanities texts.

Sets of encoding or markup tags which conform to the SGML syntax are known as SGML applications. When a text is said to be in SGML, it is important to know which SGML application and to have access to the DTD. True SGML must conform to a DTD. There are many electronic texts now in existence which claim to be SGML which do not appear to have DTDs. Others, most notably the New Oxford English Dictionary, are described as SGML-like and may not necessarily be processable by all SGML software. In some cases this may not be a problem now, but it may become so in the future as SGML becomes more widely used.

The need for standardization of markup in the humanities led to the establishment of the Text Encoding Initiative (TEI) in 1987. The TEI is sponsored by the three major text analysis computing organizations: the Association for Literary and Linguistic Computing, the Association for Computers and the Humanities, and the Association for Computational Linguistics. It has become a major international project with funding of over $1 million from North America and from the Commission of the European Communities beginning in 1988. Its objectives are to define a common encoding format for electronic texts and to provide guidelines for the interchange of electronic texts. Further information about the TEI project is available from the fileserver of the listserv <tei-l@uicvm>.

The TEI immediately made a commitment to SGML and set up four main committees to deal with different aspects of encoding electronic texts. The documentation committee defined a method of documenting electronic texts which is stored within the text as a header. This is described in more detail in the next section. The text representation committee first looked at ways of encoding the physical description and logical structure of text and identified the components and core features of basic text types. It then set up a number of work groups to look in more detail at specific areas and text types. These included character sets, hypermedia, textual criticism, language corpora, formulae and tables, verse, performance texts, and literary prose. A third committee on analysis and interpretation first devised general purpose mechanisms for encoding linguistic and other analytic interpretations which are comprehensive
enough to allow several different interpretations to be placed on a word or section of text. It then set up work groups to look at electronic dictionaries, spoken texts, and terminological data as well as the interpretation of historical material and further linguistic analysis. A fourth committee defined how best the TEI might use SGML. It prepared a kind of “house-style” for the TEI's use of SGML and proposed methods for dealing with multiple hierarchies.

The TEI Guidelines have been developed following a set of principles established at the planning meeting in 1987. The guidelines are intended for text in any kind of written or spoken language. They are intended for both scholars and librarians. The guidelines give recommendations both on what features to encode and how to encode them. The features discussed in the guidelines include both those which are explicitly marked and those which are the result of analyzing and interpreting the text. Although the TEI Guidelines include some 400 different encoding tags, very few indeed are absolutely required. The basic philosophy is “if you want to encode this feature, do it this way.” Sufficient information is provided for the TEI DTDs to be extended by users if necessary.

The TEI Guidelines are built on the assumption that virtually all texts share a common core of features, to which can be added tags for a specific discipline, text type, or application. The encoding process is seen as incremental, so that additional tags may be inserted in a text as new researchers work on the text. Almost all encoding implies some interpretation of a text and so the guidelines provide for multiple views of a text and multiple encodings for individual phenomena within a text. They also provide a means of documenting any interpretation so that a new user of the text can know why that interpretation is there.

A TEI conformant text consists of a TEI header followed by the text itself. The text has optional front and back matter. The body of a text is divided into units which, for convenience, the TEI has chosen to call divisions using the tag <div>. SGML attributes are used to identify the type of division—e.g., “chapter,” “stanza,” “act.” Within the smallest division, the basic element is a paragraph which can contain many other elements such as lists, names, dates, abbreviations, and so on.

The first draft version of the TEI Guidelines (Sperberg-McQueen & Burnard, 1990) has been distributed extensively for comment. The second draft is being made available electronically in fascicles from the listserv <tei-l@uicvm> as new chapters have been completed for publication. A cumulative print version is in preparation (Sperberg-McQueen & Burnard, 1990).
DOCUMENTING ELECTRONIC TEXTS

The source from which an electronic text has been compiled is sometimes not known or is unclear. One of the major reasons for this is that, until recently, there has been no standard way of providing this information in such a way that it does not get detached from the text or lost. Most of the large text archives held in research institutes use databases which they have developed themselves for recording information about the texts. These databases often consist of limited information of value only to themselves. Individuals who have created texts have not often even provided this information, most obviously because they themselves were fully aware of it and thus did not feel the need to record it. Many existing texts have encoding within them which is not documented and, if the exact source is not known, it may be impossible to identify, for example, what a group of percent signs in the middle of a text may mean.

One of the TEI’s major contributions is a set of proposals for documenting electronic texts so that users may know what they have and librarians will have the information they need to catalog the texts. The TEI header is believed to be the first systematic attempt to provide in-file documentation of an electronic text which conforms to the same syntax as the markup within the text. It consists of four major sections or SGML elements each of which contains further elements or subdivisions. The file description element is the most important. It contains a full bibliographic description of the electronic file which can be used for creating catalog entries or bibliographic citations. It must include a file statement which gives the title of the work and those responsible for its intellectual content, a publication statement which identifies the publication or distribution of an electronic text, and a source description, which is a bibliographic description of the source from which the electronic text was derived. Additional optional elements give information relating to one edition of a text, the approximate size of the text in whatever units are convenient, the series, if any, to which a publication belongs, and notes which provide additional descriptive information not contained in other elements.

The encoding description element provides information which the user of a text needs to know. It documents the methods and editorial principles that governed the transcription of the text, also giving the intellectual rationale for any analytic or interpretive information. Additional information which characterizes a text but does not fit easily into the other header sections is given in the profile description. This includes information about the participants in a conversation if the text is a transcript of speech as well as details
of the natural languages used in the text. The fourth section, the revision history, documents any changes made to the text and provides information which is critical for working with electronic texts in which changes are made over time and where there is a need to ensure that a particular version of a text was used.

The file description contains sufficient information for a librarian to catalog the text, with indications of its source. The encoding description contains information which anyone who uses the text needs to have. The revision history provides a means of recording updates to the text. The TEI header does not yet have elements which specifically address authentication, but it would be a simple matter to define extra elements which would contain a time stamp or other authentication codes. These may further be extended to apply to only certain SGML elements within the text, leaving the others to be modified as users exploit the text for their own purposes.

**CATALOGING ELECTRONIC TEXTS**

The only attempt to create a systematic catalog of electronic texts in the humanities using standard bibliographic procedures is the Rutgers Inventory of Machine-Readable Texts in the Humanities, which began in 1983 in response to a growing number of enquiries about the availability of electronic texts. Responsibility for the inventory was assumed by the Center for Electronic Texts in the Humanities (CETH) when it was established in the second half of 1991 at Rutgers and Princeton universities. The inventory is available now on the Research Libraries Information Network (RLIN) and should be accessible soon in other ways as well.

Research done at CETH indicates that there is little expertise in the specific problems of cataloging electronic texts. CETH has found that most of the expertise in cataloging computer files is derived from experience with either software or social science numeric data files. There seems to have been very little emphasis specifically on electronic text files in the humanities where it is necessary to know whether the text requires specific software in order to use it, or what encoding scheme it uses. The rules in Chapter 9 of the Anglo-American Cataloging Rules, second edition, 1988 Revision (AACR2R) (ALA, 1988), cover all kinds of computer files (programs, numeric data files, and so on) but are not especially suitable for electronic texts. CETH has now developed comprehensive guidelines for cataloging electronic texts in the humanities which ensure that sufficient information is provided in the catalog record to enable librarians to provide access to the text in a meaningful way (Hoogcarspel, 1993, 1994).
At present, the main source of information for updating the inventory is a survey which began in 1990 with funding from the Commission of the European Communities. A first questionnaire was distributed to some 5,000 people who were on the mailing lists of the Association for Literary and Linguistic Computing, the Association for Computers and the Humanities, the Association for Computational Linguistics, and twelve other sponsoring organizations. Recipients were asked to note which of the following types of data they hold: (1) speech (tape recordings), (2) single (individual) texts, (3) collections of texts, (4) corpora (collections of texts which have been put together on a principled basis as samples which represent a specific population—e.g., the Brown Corpus), (5) machine-readable dictionaries, and (6) computational lexica. The intention was to send a different follow-up questionnaire depending on the type of data held. CETH took over responsibility for doing the follow-up for single texts, collections of texts, and corpora.

The questionnaire reflected input from a large number of people in the field and was intended to be comprehensive in its coverage with regard to written language, spoken language, and dictionary and other lexical information. In redesigning the questionnaire, CETH took great care to ensure that bibliographic records could easily be created from the data. Examples were provided together with simple explanation. The responses have come from a variety of individuals and institutes, all of whom appear to have different procedures for documenting and cataloging their texts. None found it particularly easy to describe their material in such a way that catalog records could easily be created from it, and CETH is often finding it necessary to go back to the respondent for further information. Cataloging the backlog of existing collections is not going to be an easy task. Future material will be much easier to catalog once standard procedures for documenting the texts are adopted.

SOFTWARE ISSUES

Software issues are also central to the evaluation of electronic texts, the key question being how to provide access to the texts. An ASCII file with markup offers the most possibilities in that it is not restricted to any particular program. Scholars and libraries can use it for whatever purposes they like. However, they need to acquire software from somewhere else or write their own programs. Writing their own software gives the most flexibility, but in reality very few people are going to do that. The time investment is large and it is obviously not efficient for the person who just wants to look up a few words. It pays off only for the scholar who has very specific individual requirements.
Texts which are pre-indexed for some specific software are easier to begin with and can satisfy a good many purposes, but it is important then to look at the facilities provided by the software. Browsing normally is easy, but the real issue is how to find exactly what is wanted. The response to any search request is only as good as what is indexed. Most commonly used index-based retrieval programs cannot really handle many of the complex features of scholarly texts such as marginalia, the critical apparatus and parallel texts which were discussed earlier in this article. They work with document structures which are too simple. Accented characters and diacritics also pose problems as decisions need to be made at the time of building the index as to whether these are to be specified in a search query. Once scholars begin to use these programs seriously and gain an understanding of basic techniques, more often than not they will want to ask questions which cannot be answered by a package, even though it is technically feasible to do so. They will find that much existing software becomes limiting in terms of what it can do. There are instances where the capability of the software is driving the kinds of research which can be done and, in some cases, restricting it. Electronic texts can transform scholarship, but the needs and requirements of scholars ought to be paramount in the design and development of new software which effects these transformations.

Traditional humanities computing software, such as Micro-OCP and TACT, leaves these decisions entirely in the hands of the scholar. He or she has specific areas of research and can work with that in mind, but as more people want to work on the same texts, it makes sense not to duplicate effort and to provide these centrally. Libraries are now beginning to index texts for many different people to use, most notably with the PAT software, a retrieval program developed by Open Text Systems of Waterloo, Canada. In effect, in some ways, they are taking on responsibility for the intellectual content of the text since they have to decide what is indexed and how it can be accessed. The Text Encoding Initiative has laid the groundwork for multipurpose text after over four years of research. It is debatable whether existing software provides the capability of building indexes which can satisfy many different purposes. A good deal more research may be needed in this area, particularly in the exploitation of SGML.

Conclusion

What can a library do now? Setting up an electronic text center with adequate support is a considerable investment. The greater part of the cost will involve staff who support the facility. Ensuring that staff have adequate training in the relevant tools and techniques is essential so that they can make informed judgments. Discussion
groups, such as <etextctr@rutvml>, are intended to help librarians enter this new world. Consultation with potential users is important at all stages. Experimenting with some CD-ROM-based resources would be a good first step as well as looking at some widely available software tools. These should help libraries decide how to handle texts over the network which is so obviously the long-term future. More than anything, widespread consultation and collaboration in research will be needed in order to determine the principles and procedures for providing access to the multipurpose high-quality electronic text which will serve the needs of humanities scholarship in the next century.

REFERENCES


