The Text Encoding Initiative: Its History, Goals, and Future Development

Nancy M. Ide
Department of Computer Science
Vassar College
Poughkeepsie, New York, 12601, USA
914 452-5988; ide@cs.vassar.edu

C. M. Sperberg-McQueen
Computer Center
University of Illinois at Chicago
Chicago, Illinois, USA
312 996-2477; u35395@uicvm.uic.edu

1. Overview

Before they can be studied with the help of computers, texts must be encoded in computer-readable form. Standard data processing practice provides convenient solutions for basic text representation problems, but many texts of interest to scholarly research present difficulties not resolved by industrial standards. Therefore, over the years scholars have developed a variety of methods for representing special characters, encoding logical divisions of a text, representing analytic or interpretative information, and reducing text-critical apparatus to a single linear sequence. Because of the lack of a unified, standard format, scores of such encoding schemes were developed in the 1960's, 70's, and 80's from scratch or adapted from existing schemes. These schemes typically reflected the specialized interests of their originators and were, by and large, incompatible; the end result was that a text encoded for one purpose or piece of software often required substantial editing to be used for another purpose or with other software, if it was reusable at all. Recognizing this, the humanities computing community attempted very early to launch an effort to develop encoding standards for computer-readable texts intended for scholarly research (San Diego 1977, Pisa 1980). However, these efforts failed to generate consensus on how, or even if, such a standard should be developed, and thus they were aborted at the outset.

In November of 1987, the Association for Computers and the Humanities (ACH) convened a meeting at Vassar College in Poughkeepsie, New York, of over 30 representatives from archives, humanities computing centers, and professional organizations, to once again consider the standardization question. This group agreed not only on the need for common practice but also on a set of basic principles to guide the development of guidelines for the encoding and exchange of literary and linguistic data, now commonly referred to as the "Poughkeepsie Principles":[1]

1. The guidelines are intended to provide a standard format for data interchange in humanities research.
2. The guidelines are also intended to suggest principles for the encoding of texts in the same format.
3. The guidelines should

[1] These basic principles are expounded in various internal documents of the Text Encoding Initiative, notably TEI EDP1 and TEI EDP2, available from the University of Illinois at Chicago Computer Center.

a. define a recommended syntax for the format,
b. define a metalanguage for the description of text-encoding schemes,
c. describe the new format and representative existing schemes both in that metalanguage and in prose.

4. The guidelines should propose sets of coding conventions suited for various applications.
5. The guidelines should include a minimal set of conventions for encoding new texts in the format.
6. The guidelines are to be drafted by committees on
   a. text documentation
   b. text representation
   c. text interpretation and analysis
   d. metalanguage definition and description of existing and proposed schemes, coordinated by a steering committee of representatives of the principal sponsoring organizations.
7. Compatibility with existing standards will be maintained as far as possible.
8. A number of large text archives have agreed in principle to support the guidelines in their function as an interchange format. We encourage funding agencies to support development of tools to facilitate this interchange.
9. Conversion of existing machine-readable texts to the new format involves the translation of their conventions into the syntax of the new format. No requirements will be made for the addition of information not already coded in the texts.

The success of the Vassar conference resulted from several factors: first, at the time of the conference more was known about encoding problems and basic principles were clearer. Second, the Vassar group included a far more robust representation of key organizations and active research centers than had been gathered before. Third, the recently developed Standard Generalized Markup Language (SGML)[2] provided a tool for developing a simple, flexible, and extensible encoding scheme capable of satisfying the widely varying needs of textual research. Finally, the consensus reflected the growing urgency of the need. At earlier meetings, it was predicted that if the humanities computing community did not adopt a common practice, chaos would ensue. At the Vassar meeting, no one needed to predict chaos; it was, as several speakers observed, the status quo.

Following the Vassar conference the ACH was joined by the Association for Literary and Linguistic Computing and the Association for Computational Linguistics in driving the standards effort, thus forming the Text Encoding Initiative (TEI). The three organizations pledged to guide the effort and seek funding to support the TEI as an international, multi-lingual project to develop guidelines for the preparation and interchange of electronic texts for scholarly research.[3] Very quickly, it was recognized that the TEI's goals served not only humanities scholarship, but were critical for a broad range of applications by the language industries more generally. It has become crucial for both research and industry to ensure that any text that is created can be used and, more importantly, reused for any number of applications and for more, as yet not fully understood, purposes. Thus since its inception, the work of the TEI has achieved increasingly central importance for text-based work across disciplines and applications.

---

[2] International Organization for Standardization, *Information Processing -- Text and office systems -- Standard Generalized Markup Language (SGML)*, ISO 8879-1986 (E) ([n.p.]: International Organization for Standardization, 1986).

[3] Major support for the TEI has been provided by the U.S. National Endowment for the Humanities (NEH), Directorate XIII of the Commission of the European Communities (CEC/DG-XIII), the Andrew W. Mellon Foundation, and the Social Science and Humanities Research Council of Canada.

In the fall of 1993, the TEI issued the first full version of its *Guidelines for the Encoding and Interchange of Machine-Readable Texts.*[4] This report, which provides encoding conventions for a large range of text types and features relevant for research in language technology, the humanities, and computational linguistics, represent a major milestone: never before the TEI was it possible to achieve consensus among the research community about encoding conventions.

In developing its Guidelines, the TEI identified the encoding needs for interchange and for the varied processing and analysis needs of the research community, laid out on this basis the encoding principles demanded for a general purpose scheme, and identified key text types and features for which encoding conventions needed to be developed. In most cases there were no pre-existing encoding conventions. In almost as many cases, there had not even been a prior analysis of the required categories and features and their relations for a given text type, in the light of real and potential processing and analytic needs. The TEI motivated and accomplished the substantial intellectual task of completing this analysis for a large number of text types and provided encoding conventions based upon it.

The TEI's achievements include:

1. determination that the Standard Generalized Markup Language (SGML) is the appropriate framework for development of the Guidelines;
2. specification of restrictions on and recommendations for SGML use that best serves the needs of interchange, as well as enables maximal generality and flexibility in order to serve the widest possible range of research, development, and application needs;
3. analysis and identification of categories and features for encoding textual data, at many levels of detail;
4. specification of a set of general text structure definitions that is effective, flexible, and extensible;
5. specification of a method for in-file documentation of electronic texts that is compatible with library cataloging conventions and can be used to trace the history of the texts and thus can assist in authenticating their provenance and the modifications they have undergone;
6. specification of encoding conventions for special kinds of texts or text features:
    a. character sets
    b. language corpora
    c. general linguistics
    d. dictionaries
    e. terminological data
    f. spoken texts
    g. hypermedia
    h. literary prose
    i. verse
    j. drama
    k. historical source materials
    l. text critical apparatus

The TEI Guidelines are the result of this work. They provide encoding conventions for describing the physical and logical structure of many classes of texts, as well as features particular to a given text type or not conventionally represented in typography. They treat common text encoding problems, including intra- and inter-textual cross reference, demarcation of arbitrary text segments, alignment of parallel elements, overlapping

[4] *Guidelines for the Encoding and Interchange of Machine-readable Texts*, C.M. Sperberg-McQueen and L. Burnard, eds. (Chicago and Oxfrod, ACH-ACL-ALLC Text Encoding Initiative, 1993).

hierarchies, etc. In addition, they provide conventions for linking texts to acoustic and visual data. As such, the TEI Guidelines answer the fundamental needs of a wide range of users: researchers in the humanities, sciences, and social sciences, publishers, librarians, and those concerned generally with document retrieval and storage. They also answer many of the needs of the growing "language technology" community, which is amassing substantial multi-lingual, multi-modal corpora of spoken and written texts and lexicons in order to advance research in human language understanding, production, and translation.

In what follows, we discuss in more depth the goals of the TEI and its overall organization.

2. Rationale for an Encoding Scheme

2.1 Scope and Intent

2.1.1. Definition of "text"

The concern of the group who met at the Vassar conference was "texts intended for humanities scholarship." The range of texts included under this definition was not entirely clear; very generally, such texts can be said to include pieces of extended natural discourse, ancient or modern, in any language. We can for the most part think of texts existing in written form, although transcripts of spoken language may be included. It was not clear that concordances, word lists, results of linguistic surveys, and other items lacking the interrelational coherence and co-referentiality of continuous discourse meet the implicit criteria for textuality assumed at the Vassar conference. Dictionaries, which are not composed of pieces of continuous text but whose co-referentiality is extensive, clearly exist on a borderline and were eventually taken by the Vassar group to be included under the rubric "text."

Whatever the boundaries, the needs of humanities research were not fully addressed by schemes such as the Association of American Publishers' standard for encoding materials for eventual typesetting.[5] Computer-readable texts intended for research occasionally use mark-up to describe potential physical layout, but typically include very different types of information, such as bibliographic information, physical description of an existing form or forms of the text (with no intention for reproduction in this form), information concerning the logical structure, and interpretive or analytic information concerning semantic or linguistic elements within the text.

Over time the range of text types to be covered by the TEI and the community it intended to serve was broadened, as it became clear that the needs of any textual research within or outside the humanities, as well as the growing number of researchers and users of text in industry, were largely overlapping. The growing diversity of applications for electronic texts includes not only humanities research but also natural language processing (machine translation, language understanding, etc.), information retrieval, hypertext, and electronic

---

[5] Association of American Publishers, *Reference Manual on Electronic Manuscript Preparation and Markup*, The Association of American Publishers Electronic Manuscript Series (Washington, D.C.: Association of American Publishers, 1986). Association of American Publishers, *Author's Guide to Electronic Manuscript Preparation and Markup*, The Association of American Publishers Electronic Manuscript Series (Washington, D.C.: Association of American Publishers, 1986). Association of American Publishers, *Markup of Mathematical Formulas*, Association of American Publishers Electronic Manuscript Series (Washington, D.C.: Association of American Publishers, 1986). Association of American Publishers, *Markup of Tabular Material*, The Association of American Publishers Electronic Manuscript Series (Washington, D.C.: Association of American Publishers, 1986).

publishing. An early TEI emphasis on encoding linguistic information, such as morphology and syntax, reflects the recognition that such encoding was fundamental to scholars and researchers across a broad range of disciplines and applications.

2.1.2. Guidelines vs. standards

The TEI made an early commitment to provide *guidelines*, rather than a *standard*, for encoding literary and linguistic materials. The goal was to specify encoding conventions that are coherent, easy to use, relatively comprehensive, and which provide ample mechanisms for user-defined extensions in order to suit individual needs. It was recognized that whether or not the TEI scheme is constituted as an established standard, users are obviously free to adopt it or to devise their own scheme; but it was hoped from the outset that if the TEI Guidelines meet the criteria outlined in the Poughkeepise Principles, the need to devise an independent scheme would be obviated in many instances and the long-recognized need for uniformity in the encoding of computer-readable texts would be realized.

The participants in the Vassar conference included several representatives of major archives, who were both anxious to promote the idea of a common format for text encoding, but at the same time hesitant to convert their existing holdings into a new format. From this developed the notion of the TEI Guidelines as first of all a format for data *interchange* as opposed to data *storage*. This emphasis enabled archives to retain their data in internally-developed formats keyed to locally developed software, and required only that they convert to the TEI format for interchange. Users would acquire texts in a single, familiar format potentially compatible with commercially available analytic software, and the archives would need only to develop programs to convert texts to and from a single, common format into their locally defined encoding system.

The concern over conversion between formats led to the third Poughkeepsie Principle, which stressed that the TEI should develop a metalanguage for describing encoding schemes. The idea behind this principle was to provide a formal description of the mapping among schemes, in order to facilitate the conversion between schemes envisaged at the Vassar conference. Among the Poughkeepsie Principles, this is the only one which was explicitly dropped later on in the project. There are several reasons for this: first, the concern over the difficulty of mapping between encoding schemes was early diminished as it was recognized that this mapping was in nearly every case straightforward--in part because the SGML-based encoding conventions developed within the TEI have been devised with a maximum eye toward flexibility and generality. Second, since the Vassar conference SGML has gained much wider acceptance throughout the research and industrial communities, and many archives are adopting it in any case fro both internal and external use. Finally, no one anticipated in 1987 the volume of new texts which would subsequently be encoded and made available to the research community. Hundreds of millions of words of newly-encoded text in many languages are becoming available, most of which are encoded with at least an eye toward SGML and the TEI.

The Guidelines were of course also intended to provide recommendations for newly-encoded texts---specifically, to assist scholars and research centers with no commitment to an existing encoding format in deciding both *what* textual features to encode and *how* to encode them. Recognition of this goal led to three desiderata for the Guidelines :

1. the Guidelines should specify a *recommended* minimum set of tags to be included in every newly-encoded text, including descriptive and bibliographic information as well as information concerning the encoding itself.

2. the Guidelines should define the textual features relevant to specific disciplines or text types, and define tag sets to enable marking these features within a text.
3. because the varieties and needs of both textual materials and research defy exhaustive classification, the Guidelines should include a mechanism to enable users to extend the scheme.

2.1.3. Polytheoreticity

The task of defining relevant textual features within specific disciplines in many cases raises questions concerning the theoretical or practical orientation represented by tag sets. At one extreme, tags developed for application in a certain field might represent features defined by a single theory---for example, generalized phrase structure grammar in the field of linguistics---with no consideration of other major theories in the field or of the relation among sets of features defined by competing theories. At the other extreme, consensus among competing theories or systems might be aimed for, in order to develop a theoretically neutral or "polytheoretical" tag set for such applications. In the latter case, achieving consensus could involve considerable research or may be impossible, within the current theoretical climate, to achieve.

Consensus is an appealing goal and has been attempted in cases where it was felt possible to achieve it readily. In cases where consensus was clearly problematic, a variety of alternative approaches to the development of tag sets was adopted. The most straightforward involved formal specification of a separate set of features and the structure of relations among them, for each major competing theory or system in current practice. In other cases (most notably linguistic annotation--see Langendoen and Simons in this issue) methods for formally defining alternate meanings for tags within a tag set have been provided, so that one set of tags serves for several alternate theories or systems and is interpreted in any given application according to explicit specifications by the user. A third approach was to determine a minimum set of features which includes as a subset the appropriate features of each competing theory or system. Determining tag sets for each theory or system independently was a useful prelude to the other approaches, and constituted a first step in every case.[6]

2.2 Syntactic Issues

2.2.1. SGML

The participants in the Vassar conference agreed unanimously that the TEI Guidelines should not only specify *what* the user should or could encode in a computer-readable text for various applications in humanities scholarship, but also *how* these features should be encoded---that is, a concrete syntax for the recommended and suggested tags. No final decision about the syntactic basis for the new encoding scheme was made at the conference, but it was agreed that if possible, the syntax should be borrowed from an existing scheme, be relatively simple to use, and be capable of expressing the fine distinctions and occasionally complex overlapping hierarchical structures required in textual data. In addition, the conference mandated that the syntax of the Guidelines should be designed to ensure device independence within the data stream. A third goal was compatability with existing standards. Consequently, the Standard Generalized Markup Language (SGML) was seen as the most likely candidate to provide a syntactic basis for the Guidelines.

---

[6] For a fuller treatment of the TEI approach to polytheoreticity, see "Theoretical Stance and Resolution of Theory Conflict," TEI internal document TEI EDP3, available from the University of Illinois at Chicago Computer Center.

SGML, which is a syntactic framework for developing tags sets rather than a tag set itself, was early adopted as the basis of the Association of American Publishers' standard for electronic manuscript markup, which had wide acceptance in the scholarly community.[7] A survey of encoding problems at Queens University in 1986 concluded that SGML offers a better basis for research-oriented text encoding than other schemes,[8] in large part because of its orientation toward *descriptive* markup (markup which describes function rather than form--e.g., "emphasis" or "foreign word" rather than "italics").[9] Since 1987, SGML has become widely adopted by government, industry, and academic groups world-wide. Thus the TEI Guidelines, by adopting SGML, have achieved de facto compatability with a large number of other encoding schemes in addition to that of the AAP.

2.2.2. Software and Application Independence

The TEI scheme was from the outset intended to be hardware-, software-, and application-independent.

Software independence has meant that the current capabilities and limitations of SGML-processing software have not played a determining role in choices made in the design of the TEI scheme. The TEI Guidelines are intended to serve for many years to come, and it would be foolish to design them to accomodate existing software. <<michael do you have something more to say here>>>

Application-independence has meant that the TEI has not, in particular, been driven by the notion of electronic text as a stage in the production of paper documents. Like the publishing industry, the academic community is rapidly coming to realise that its stock in hand is not words on the page, but information, independent of its physical realization. Thus in its design the TEI has also embraced a view of electronic text as an end in itself, whether as a research database or a component in non-paper publications.

Application-independence, coupled with the TEI's commitment to serve the full range of research interests, also means accomodating different views of a text. In different contexts, texts may be regarded as

* physical objects (volumes or loose leaves of paper, parchment, or papyrus with ink in specific places; or acoustic signals occurring at a particular time and place; or clay tablets or stones with a three-dimensional writing surface)
* typographic objects (series of characters in specific fonts, laid out and justified in a particular style)
* linguistic objects (series of graphemes or phonemes, or at a higher level series of morphemes or lexical items or phrases or sentences)
* formal objects (series of stanzas, cantos, acts, chapters, sections, etc., in turn subdivided into smaller formal units)
* rhetorical objects (series or hierarchies of speech acts, rhetorical figures, tropes)

---

[7] See note 1, above.

[8] Cheryl A. Fraser, ``An Encoding Standard for Literary Documents,'' M.S. Thesis (Queen's University, Ontario), 1986. The work was performed under the direction of David T. Barnard.

[9] James H. Coombs, Allen H. Renear, and Steven J. DeRose, ``Markup Systems and the Future of Scholarly Text Processing,'' *Communications of the Association for Computing Machinery*, 30:11 (Nov. 1987), pp. 933-47.

* propositional objects (referring to specific persons, things, places, and events, real or imaginary, in ways subject to paraphrase and abstract representation)
* historical and cultural objects (with strands and layers of witness-es to the textual transmission, interpretation, re-interpretation, and commentary)

The TEI Guidelines define a general-purpose encoding scheme which enables encoding any of these views. Further, it enables the simultaneous encoding of multiple views, which is important for both research and industrial text applications. For example, the scholar reconstructing the lexicon of an ancient language from surviving parchment fragments or an industrial application for document translation must constantly switch between the levels of physical and linguistic description; the historian or anthropologist testing a theory of social interactions or customs by an investigation of textual records relating to them must switch between linguistic and propositional perspectives. No absolute recommendation to embody one specific view of text can apply to all texts and all approaches to it. The TEI scheme therefore provides multiple ways to encode the same feature in many cases.[10]

3. Organization of the Project

3.1. General Organization

A small central organization coordinates the work of the TEI. Two representatives from each of the three sponsoring organizations (ACH, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing) form a Steering Committee which oversees the project. An editor in chief and an associate editor have been responsible for the centralized work and for elaborating the basic design produced at the Vassar meeting.

To help ensure that the TEI Guidelines reflect the needs of scholarly research, an Advisory Board representing professional groups for literary, linguistic, and historical research and teaching as well as computing, library, and publishing organizations is responsible for approving the content of the Guidelines at various stages in their development. These organizations are:

Modern Language Association
Association for History and Computing
American Historical Association
Association for Documentary Editing
American Philological Association
American Philospohical Association
Association Internationale de Linguistique Appliquee
Linguistic Society of America
American Society for Information Science
Association Internationale du Bible Informatisé

[10]At the outset of the TEI, it was not clear that SGML could adequately handle multiple overlapping hierarchies which often result from the encoding of multiple views (e.g., *canto, stanza, line* on the one hand and *poem, sentence, word* on the other). It had been shown that overlapping hierarchies can be defined over a text but not that they could be processed simultaneously (see Barnard, David, et al. "Using SGML to Maintain Multiple Structures in Documents," External Technical Report, ISSN-0836-0227-87-204, (Kingston, Ont.: Queen's University Department of Computing & Information Science, Dec. 1987), and Barnard, David, et al. "SGML-Based Markup for Literary Texts: Two Problems and Some Solutions." (forthcoming)). This problem received considerable attention within the TEI, the results of which are described in Price, "Hierarchical Encoding of Text: Technical Problems and SGML Solutions", in this issue.

3.2. Committees

Most standards-development efforts are voluntary, and the effort to develop the TEI Guidelines has been more voluntary than most. From the outset it has been clear that the Guidelines must reflect the consensus of those interested and at the same time take into account the special needs and special desires of everyone who is to use them. It was therefore important to involve many different people, with differing areas of expertise including discipline-specific expertise as well as technical and SGML-specific expertise, in the design process. The success of the TEI is in particular the result of the donation of time and expertise by the many members of the wider research community who served on the TEI's Committees and Working Groups.

Four committees were initially responsible for producing appropriate sections of the Guidelines; in most cases, several specialist Working Groups within these committees worked on developing schemes for specific areas.

3.2.1. Committee on Text Documentation

The committee on text documentation was given primary responsibility for the "prolog" or "label" section of a TEI-conformant document. They defined tags for the information *about the text* which encoders are encouraged, or may wish, to provide. This meta-textual information falls into four classes:

1. identification of the text itself, in sufficient detail that the user can locate either the original copy text or some other edition of the text encoded);
2. identification of the encoding itself sufficient to allow library cataloguers or text archivists to catalog the files;
3. description of features of interest to archivists and their borrowers;
4. declaration of special features of the encoding, so that programs can process the text properly. The text documentation committee must provide a location for this information and prevent conflicts among various declarations, but the syntax and content of the declarations will be determined by the other committees.

This committee was competent in bibliographic description and archive management. Fortunately, there existed when they began their work a well-developed discipline for bibliographic description, both for texts on paper and for machine-readable data files. The committee worked from the International Standard Bibliographic Descriptions for most text types, supplementing them from other sources were necessary. Experienced data archivists recommended tags for the kinds of information (apart from a good bibliographic description) that they find most useful and important in dealing with their borrowers.

The result of this committee's work was the TEI header, described in Giordano, and treated extensively in Dunlop, both in this issue.

3.2.2 Committee on Text Representation

The committee on text representation provided for the adequate representation of printed or manuscript versions of the text. This includes:

1. the "physical" description of the copy text
2. the "logical" description of the text, with tags for the textual features conventionally represented by typography in a printed edition (whether present in the copy text or not), including:
    a. special characters, symbols, and non-Latin alphabets

b. the structural hierarchy of the text (e.g. book, chapter, verse)
c. common typographically realized text features (e.g. emphasis, quotation, tabular layout, etc.)
d. less common or special text features (e.g. notes, marginalia, commentary, parallel texts editorial emendations, and critical apparatus)

Scholars have already devised solutions for most of the problems faced by this committee, notably character set issues (see Gaylord and Tutiya in this issue) and the delineation of text structure (see for instance Johannson, Lavagnino and Mylonas, and Chisholm and Robey in this issue). With these and the AAP tag set as a starting point, the committee formulated a single coherent solution, including recommendations for the common cases, procedures for documenting deviations if the recommendations are not followed, and procedures for declaring character sets or structural tags in the cases to which the recommendations do not apply.

In the first phase of this committee's work, the focus was on the logical description of the text, reserving detailed physical description for later phases. The tag set proposed is intended to be adequate to the fundamental needs of unillustrated literary texts (poetry, plays, novels and short stories) in both critical and popular editions. Codes are provided only for alphabetic languages; methods of encoding multidirectional text will be treated in future version of the Guidelines. This tag set was later extended to handle problems presented in less common text types, more general cases of reference works, and more complex tabular and mathematical material.

>>>More complex types of apparatus and commentary will be handled at this stage, based on experiences with the initial set of tags. Multidirectional text will be handled at some length, and provision for declaring the encoding method used will be made. An attempt will be made to provide guidance for the encoding of every language in which computer-assisted work is known to be underway in Europe or North America. In addition, a detailed physical description of the text carrier adequate for the types of codicological and typographical research currently pursued by machine will be provided.<<<michael has to revise this

3.2.3. Committee on Text Analysis and Interpretation

This committee was charged with providing tags for textual features not conventionally represented typographically in a text. For several scholarly fields and research areas, it provided specific tag sets for recording textual features (objective or subjective, given or achieved as the result of analysis or study) of interest to researchers in that field.

The work of this committee can be broken down into the problems presented by various types of textual study:

1. problems common to many fields (e.g. intratextual and intertextual cross reference, demarcation of arbitrary text segments with pointer to commentary or other related material, tags for indexing text items or segments with arbitrary terms of interest to the scholar, etc.);[11]
2. linguistic analyses (e.g. tags for corpora, dictionaries, syntax, morphology, and lexical analysis);[12]

[11] See DeRose and Durand in this issue.

[12] See Dunlop, Ide and Véronis, and Langendoen and Simons in this issue.

3. literary study (e.g. tags for thematic study, identification of allusions, marking for traditional narrative materials like myths, meter, prosody, and the structural analysis of narrative).[13]

As noted earlier, the committee was forced to decide, within any field, whether to provide separate or overlapping tag sets for any competing theoretical approaches, to attempt a union of the various sets of textual features they tag, to delimit the areas of difference as they affect the tagging of the text and allow the encoder to declare the use of specific positions or practices in the areas of difference, or to unify the various positions in a theory-neutral or poly-theoretical tag set.

3.2.4. Committee on Metalanguage Issues

The committee on metalanguage issues was responsible for providing a syntax for the tag set of the Guidelines.

The syntax of the international standard SGML was adopted as the basis for all work on the tag set of the Guidelines themselves. The committee also legislated on the features of SGML which would be generally adopted by the TEI in its recommendations for an interchange format, as well as on specific syntactic solutions to problems encountered within various work groups (see Price, "Hierarchical Encoding of Text: Technical Problems and SGML Solutions", in this issue for a discussion of some of these).

3.3. Affiliated Projects

It was recognized from the outset that the Guidelines will be successful only if they prove useful to those who are actually encoding texts. While the working committees will encode texts and text fragments in the course of their work, it was seen necessary to try the Guidelines out on larger bodies of material if possible. This required the cooperation of current encoders of significant bodies of material.

The TEI established liaison with several large encoding projects, including:

>> include list of ap's

For each, the TEI provided drafts of portions of the Guidelines (inclduing drafts and internal copies) as soon as they were available, and provided consulting on the TEI scheme.

In exchange, the TEI requested that these affiliated projects review TEI materials, provide feedback on their utility and clarity, report all problems they encounter in applying the encoding scheme, give permission to use extracts of their work as examples in our documentation, and (as appropriate) serve on working committees.

3.4. End Products

In November 1990 the TEI produced a first draft of its Guidelines, called TEI P1 ("proposal 1"),[14] which was subsequently widely distributed and discussed. Starting in

---

[13] For meter, see Chisholm and Robey in this issue. Other areas of literary analysis are not treated in TEI P3, due to serious problems arising from so far unresolvable differences of opinion concerning fundamental questions of the appropriateness of marking such items at all within the literary critical community. Work in this area will continue within the TEI.

April 1992, the TEI began publishing in electronic fascicles various chapters of what would become its next draft, TEI P2.[15]

In fall, 1993, the TEI published the most comprehensive draft to date, TEI P3, intended to be its first full proposal for text encoding Guidelines. The Guidelines describe methods of text encoding corresponding, in its level of technical detail, to a reference manual for a major software package. This document includes formal SGML document type definitions (DTDs) specifying the syntax and usage of the tag sets formulated by the Guidelines. The DTDs are also available in machine-readable form.[16]

4. Future of the project

The TEI has achieved a major milestone in establishing an intellectual foundation for text encoding and a set of encoding conventions substantial enough to serve the fundamental needs of most encoding projects, both large and small. However, much of this development has necessarily taken place in advance of experience. It is essential to continue the work if the TEI by extending the Guidelines more broadly and providing materials and facilities for user support. In addition, now that the core of a coherent set of encoding practices has been established, it is critical to provide for extensive evaluation and testing in large-scale use, and to implement mechanisms for continued extension and modification of the Guidelines in response.

The best way to promote a standard is to develop resources and software that embody it. Therefore, the primary focus of the TEI must shift to the wide-spread and large-scale implementation of the Guidelines. Actual use of the Guidelines will become the major force driving the development of extensions and modifications to it. Activity within the TEI will focus on user support, instruction, consulting, etc. One of the primary roles of the TEI will be to form a liaison with and provide consultancy for users, as appropriate, to ensure compatibility with the Guidelines as they currently exist, and to incorporate the results eventually into future versions. Another central concern of this phase will be systematic evaluation and review, again accomplished on the basis of actual experience using the Guidelines, the results of which will also guide the further development of the Guidelines.

Extension of the Guidelines will continue, to incorporate modifications, revisions, and extensions suggested or required on the basis of user responses; provide refinements and further developments of chapters in the current version; and form or encourage work groups for areas that have only been outlined, for example, physical description (manuscripts, papyri, inscriptions, etc.), literary analysis and interpretation, alignment mechanisms for multilingual corpora and for coordinating speech with speech transcriptions, multimedia processing, etc.

---

[14] *Guidelines for the Encoding and Interchange of Machine-readable Texts*, C.M. Sperberg-McQueen and L. Burnard, eds. (Chicago and Oxfrod, ACH-ACL-ALLC Text Encoding Initiative, 1990).

[15] These fascicles are available via ftp from... Information about these documents, as well as other TEI information, is available through the LISTSERV list TEI-L@UICVM.UIC.EDU.

[16] TEI P3 is available from ...

5. Conclusion

The TEI is satisfying a need recognized by the research community, by industry, and by government funding agencies -- in North America, in Europe, and in Japan.[17] The TEI is well established internationally, and its role in international coordination is critical for the future development of standards for tagging electronic texts. The TEI has established or will establish relations with a variety of related efforts and proejcts, including standardization efforts (ISO, HYTIME, EAGLES), text collections (LDC, DCI, ECI, NERC, CLR, etc.), evaluation and development efforts (EAGLES), text access efforts (CNI, CETH), and software developers (commercial SGML discipline-specific academic and research efforts, the Text Software Initiative). Through these collaborations and through the continued contributions of the research community to its further elaboration, the TEI scheme should provide the basis of the uniform encoding scheme envisaged at Vassar.

The TEI scheme is not complete and it will demand more years of effort to answer every text encoding need. Nonethless, the considerable achievement of the TEI to date cannot be ignored. Its million dollars of funding and five year duration are minimal in comparison with other projects with much smaller scope and intent.