Computers and the Humanities **29:** 5–15, 1995.

The TEI: History, Goals, and Future *

Nancy M. Ide
Department of Computer Science, Vassar College, Poughkeepsie, New York, 12601, U.S.A.
e-mail: ide@cs.vassar.edu

and

C.M. Sperberg-McQueen
Computer Center, University of Illinois at Chicago, Chicago, Illinois 60680, U.S.A.
e-mail: u35395@uicvm.uic.edu

Key words: TEI, electronic texts, text encoding, encoding standards, SGML, tagging

Abstract

This paper traces the history of the Text Encoding Initiative, through the Vassar Conference and the Poughkeepsie Principles to the publication, in May 1994, of the *Guidelines for the Electronic Text Encoding and Interchange.* The authors explain the types of questions that were raised, the attempts made to resolve them, the TEI project's aims, the general organization of the TEI committees, and they discuss the project's future.

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 from scratch or adapted from existing schemes in the 1960s, '70s, and '80s. 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 efforts 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 whether, 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 consider once again the standardization question.[1] 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":[2]

* Nancy Ide is Associate Professor and chair of Computer Science at Vassar College, and Visiting Researcher at CNRS. She is president of the Association for Computers and the Humanities and chair of the Steering Committee of the Text Encoding Initiative. C. M. Sperberg-McQueen is a Senior Research Programmer at the academic computer center of the University of Illinois at Chicago; his interests include medieval Germanic languages and literatures and the theory of electronic text markup. Since 1988 he has been editor in chief of the ACH/ACL/ALLC Text Encoding Initiative.

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
   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 had several sources. First, at the time of the conference more was known about encoding problems and basic principles were clearer than at the time of the earlier efforts already mentioned. 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)[3] 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.[4] 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, including applications which have not yet been imagined or developed. Thus, since its inception, the work of the TEI has achieved increasing importance for text-based work across disciplines and applications.

In May 1994, the TEI issued the first full version of its *Guidelines for Electronic Text Encoding and Interchange*.[5] 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, represents a major milestone: never before the TEI was it possible to achieve such broad 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 any usable 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 serve the needs of interchange among unlike systems, while retaining the generality and flexibility that make SGML suitable for a broad range of needs in research, development, and applications;
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, and overlapping hierarchies. 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.[6] 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 (not necessarily with the intention of reproducing it 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 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 formulating its encoding scheme as a set of *guidelines*, rather than as a *standard*, for encoding literary and linguistic materials. This reflects first of all a commitment to preserving the intellectual autonomy of researchers who encode texts electronically: by constraining the allowable forms and intellectual content of electronic texts, a strictly normative standard would risk constraining their ability to reflect the particular intellectual commitments of individual researchers, and thus in some ways the types of intellectual work which can conveniently be undertaken. A wholly permissive encoding scheme, on the other hand, risks failing to provide a usable basis for the sharing and reuse of expensive textual resources. It also encourages pointless variation in encoding methods arising not from different intellectual approaches to textual research, but from merely random differences in the use of the scheme.

The TEI has attempted to steer a middle course:

1. Rather than generalities and non-binding advice on the use of SGML in text encoding, the TEI has developed a specific SGML document type definition, which allows the encoding of a specific (and thus finite) set of textual features. Explicitly normative language is used to define this DTD, with systematic gradations in the modal verbs used to describe the use of each SGML element and attribute. Distinctions among required, recommended, and optional practice are common in standards documents; the TEI Guidelines differ from most such documents in the much smaller relative weight of its requirements, and the much larger weight attached to the definition of purely optional practices.
2. In the body of the Guidelines, alternative methods of recording specific textual features are provided wherever the choice is felt to reflect important aspects of the encoder's understanding of the text, or the kind of processing to be performed; variations which do not reflect such important issues of opinion have been suppressed as far as possible, with one significant exception. Common applications of some extremely general SGML elements have been given shorter expressions which are, by definition, exactly synonymous with a longer expression using a more general element; examples of such "syntactic sugar" include the provision of specialized elements for indications of grammatical gender, number, inflectional type, and the like, which are specializations of the more general (and hence at once more flexible and more verbose) element for grammatical information of any type.
3. An explicit and precise definition of *TEI conformance* is given, which should allow clear distinctions to be made between conforming and non-conforming uses of the TEI scheme.
4. Explicit mechanisms for modifying and extending the encoding scheme are defined, and the definition of conformance guarantees that documents using such extensions can be strictly TEI-conformant. Thus, even if the TEI scheme is adopted as normative in some contexts, the intellectual freedom of researchers to modify the scheme however they see fit remains guaranteed. This does not prevent funding agencies or other bodies from imposing requirements beyond those of TEI conformance, nor software developers from electing to support only the subset of TEI-conformant documents which use no extensions – but it does ensure that the notion of TEI conformance cannot itself be used to enforce intellectual conformity.
5. Finally, since the TEI scheme has been developed not by any standards agency, funding body, or other authority, but by and for the research community, researchers who do not find it useful remain free to ignore it entirely and to develop their own encoding scheme whenever they see fit.

The description of the TEI scheme as a set of guidelines rather than a standard is one way of attempting to avoid imposing unwelcome and counterproductive burdens on the text computing community. A second is the explicit distinction, reflected in the Poughkeepsie Principles, between the use of the Guidelines for local storage and processing and their use in *interchange*.

The participants at the Vassar conference included several representatives of major archives, who were anxious to promote the idea of a common format for text encoding. At the same time they were loath to incur the costs of converting their existing holdings into a new format, or to abandon their substantial investments in local expertise, software, and systems built around other encoding schemes, often schemes developed for the individual archive. Such archives have no need to adopt the TEI scheme for local storage and processing, since they already have schemes satisfactory for those

purposes, but they may still find it useful as an *interchange format*. Many existing archives pursue an active program of exchange of data with other archives in their field; such exchanges, however, typically require the reformatting of texts from one archive's format to the other's. Direct translation from one format to the other requires a new pair of translation programs for each pair of archives exchanging material – ninety such programs, for ten archives all exchanging data. Indirect translation by means of a common interchange format like the TEI requires fewer programs, only two per archive (for ten distinct archival formats, twenty programs: ten to import data from the TEI interchange format to the archival formats, and ten to export data from the archival format into the interchange format). Archives also supply texts to individual users; distributing archival texts in a single, familiar, publicly documented format supported by commercially available SGML software would represent a significant convenience for the individual and a serious reduction in the consulting load on the archive.

The concern over the specter of enforced retrospective conversion led first to the specification, in the ninth Poughkeepsie Principle, that such conversion must not require the addition of new information to existing material, which has been fulfilled by keeping the set of required elements very small. It also led to the recommendation, in the third Principle, that the TEI should develop a formal metalanguage for the description of existing encoding schemes and their mapping into the TEI scheme. This recommendation – alone among those of the Vassar meeting – was explicitly abandoned by the TEI later on in the project, for several reasons. First, anxiety over the translation of existing schemes into TEI markup subsided as the TEI scheme took shape and it became clear that for the majority of existing research texts, the mapping was normally straighforward. Second, since the Vassar conference SGML has gained much wider acceptance in both the research and industrial communities, and many archives and projects are adopting SGML for both internal and external use. Finally, the volume of new texts being encoded since 1987 has shifted the balance of concern significantly away from the problems of converting pre-1987 legacy data. It was explicitly foreseen at the Vassar Conference that the then existing resources would ultimately be only a small proportion of the available electronic texts, but the speed with which this prophecy was fulfilled has surprised some of the prophets themselves. Hundreds of millions of words of newly encoded text, in many languages, are becoming available now, a large proportion of them – perhaps even most – 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* text 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*

It is easy to define the textual features relevant to a specific discipline, when there is unanimity within the discipline as to what they are and how they behave. In general, however, no such unanimity prevails, and different theories of the subject contend within a discipline for adherents, each postulating different kinds of features in the text and different behavior of those features. In linguistics, for example, different schools postulate different parts of speech – and some assign no meaning at all to the category "parts of speech". Other disciplines are divided less visibly by overt differences of theory than by differences of opinion on the best approach to some particular practical question.

Such diversity of opinion poses a difficult problem for the definition of a general-purpose encoding scheme suitable for research work, for which different solutions are possible. At one extreme, the encoding scheme might adopt the viewpoint of a specific single theory (e.g. the one apparently commanding widest belief) – a tag set for encoding syntactic structure, for example, might use the theory of generalized phrase structure grammar, with no consideration of the major theories in the field. Needless to say, those who subscribe to the theory in question will find the resulting tag set useful, while those who find the theory uncon-

genial will find the tag set a positive hindrance in their work.

Other approaches are also possible. The tag set may pluralistically allow the user to adopt any of several competing theories. It may eclectically mix the concepts of many theories (thus allowing, at least in principle, encodings reflecting hybrid mixes of theories which might be disowned by purists on either side). In the extreme case, an encoding scheme might formulate a common ground among the competing theories, a theoretically neutral or polytheoretical tag set which enables a text to be encoded according to any one of several viewpoints, and which captures any commonalities among the theories, in the same way that a database schema captures the commonalities among the various views of the data and allows each view to be derived systematically from the underlying database. However, the formulation of such a polytheoretical view can often involve considerable research, and may prove impossible within the current theoretical climate.

In the TEI, work groups attempted to formulate, as well as they could, the consensus of the particular discipline in question. Where consensus was clearly not to be had, a number of alternative approaches were adopted. In some cases (e.g. in the encoding of dictionary entries and term-bank entries), two different encoding schemes have been formulated (in these cases, one reflecting a relatively tightly prescribed entry structure, the other reflecting a much looser structure). In other cases (e.g. the treatment of quotation marks or end-of-line hyphenation), a number of alternatives have been formally defined, and the choice of alternatives may be declared formally in the TEI header. In the case of linguistic annotation, a feature structure notation and feature system declaration have been developed, which allow the theoretical postulates governing a particular set of annotations to be documented formally and explicitly.[7]

2.2 *Syntactic issues*

2.2.1 *SGML*

The participants in the Vassar conference agreed unanimously that the TEI Guidelines should specify 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.

SGML, which is a meta-language for the specification of tag sets rather than a tag set itself, was early adopted as the basis for the Electronic Manuscript Project of the Association of American Publishers.[8] 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,[9] 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").[10] Since 1987, SGML has been 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 solely to accommodate existing software.

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 realize 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 inter-

ests, 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);
- 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 witnesses 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.[11]

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 Appliquée
- Linguistic Society of America
- American Society for Information Science
- Association Internationale Bible et Informatique

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 "header" 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. Other work groups also forwarded suggestions to this committee, when their work indicated the need for items in the header.

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

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[12] and the delineation of text structure.[13] With these and the AAP tag set as a starting point, the committee formulated a single coherent solution, including recommendations for 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. The 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 for text critical work have since been covered based on experiences with the initial set of tags.[14] In addition, a mechanism which effectively constitutes an extension of SGML, called the Writing System Declaration, has been developed to enable the encoding of every language in which computer-assisted work is known to be underway in Europe or North America.[15]

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 pointers to commentary or other related material, tags for indexing text items or segments with arbitrary terms of interest to the scholar, etc.);[16]
2. linguistic analyses (e.g. tags for corpora, dictionaries, syntax, morphology, and lexical analysis);[17]
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).[18]

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. Similarly, for each area it was necessary to decide upon the degree of generality of the scheme provided – that is, whether the recommended encoding strategy would be general enough to enable representation of even the most exotic of structures for a given text type, or whether it would provide a more constrained but more exact scheme describing the structure of the majority of texts of that type but possibly not accomodating the few extreme variants. This problem and one committee's solution are treated extensively in Ide and Véronis, in this volume.

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 Barnard *et al.* in this volume 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 over the course of the development of the Guidelines. The TEI provided drafts of portions of the Guidelines (including drafts and internal copies) to each affiliated project 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.

The degree of involvement naturally varied among the various projects; among the affiliated projects most active during the development of the Guidelines are:

- Nietzsche Nachlaz Project (Stanford, later Dartmouth)
- Brown University Women Writers Project
- Vassar/CNRS Electronic Dictionaries Project
- Perseus Project (Harvard University)
- Middleton Edition Project (Brandeis University)
- Global Jewish Database (Bar-Ilan)
- Leiden Armenian Database
- Stockholm-Umea Corpus of Modern Swedish
- British National Corpus
- Network of European Corpora

3.4 *End products*

In July, 1990, the TEI produced the first public draft of its Guidelines, and made them freely available under the document number TEI P1 ("proposal 1"); a corrected reprint (version 1.1) was distributed in November, 1990. Starting in April 1992, the TEI began publishing in electronic fascicles various chapters of the second public draft, TEI P2.

The initial development phase of the TEI ended in early 1994 when the TEI formally published its Guidelines for electronic text encoding and interchange as

document TEI P3. The Guidelines describe methods of text encoding corresponding, in their level of technical detail, to a reference manual for a major software package. TEI P3 also includes the formal SGML declarations for the tag sets it describes. It is available in both paper and electronic form.[19]

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 widespread 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; to provide refinements and further developments of chapters in the current version; and to form or to 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.

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.[20] 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 is working to establish relations with a variety of related efforts and projects, including standardization efforts (e.g., ISO, HyTime, the Expert Advisory Group on Language Engineering Standards [EAGLES]), text collections (e.g., the ACL Data Collection Initiative, the European Corpus Initiative, the Network of European Research Corpora, the Consortium for Lexical Research), evaluation and development efforts (EAGLES), text access efforts (the Coalition for Networked Information, the Center for Electronic Texts in the Humanities), 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.

Notes

[1] This meeting was funded by the U.S. National Endowment for the Humanities.

[2] These basic principles are expounded in various internal documents of the Text Encoding Initiative, notably TEI EDP1 and TEI EDP2, available from the TEI central office at the academic computer center of the University of Illinois at Chicago.

[3] ISO (International Organization for Standardization). ISO 8879:1986 / A1: 1988 (E). *Information Processing–Text and Office Systems–Standard Generalized Markup Language (SGML), Amendment 1*. Published 1988-07-01. [Geneva]: International Organization for Standardization, 1988.

[4] Major support for the TEI has been provided by the U.S. National Endowment for the Humanities (NEH), an independent federal agency; Directorate General 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.

[5] Association for Computers and the Humanities (ACH), Association for Computational Linguistics (ACL), and Association for Literary and Linguistic Computing, *Guidelines for Electronic Text Encoding and Interchange* (TEI P3), ed. C. M. Sperberg-

McQueen and Lou Burnard (Chicago, Oxford: Text Encoding Initiative, 1994).

[6] 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). This encoding scheme was adopted by the National Information Standards Organization as standard ANSI/NISO Z39.59; a revised version has recently been adopted as ISO 12083.

[7] See Langendoen and Simons, in this volume.

[8] See note 1, above.

[9] Cheryl A. Fraser, "An Encoding Standard for Literary Documents", M.S. thesis, Queen's University, Ontario, 1986. See also D.T. Barnard, C.A. Fraser and G.M. Logan, "Generalized Markup for Literary Texts", *Literary and Linguistic Computing*, 3, 1 (1988), 26–31.

[10] 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), 933–47.

[11] 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 D. Barnard, R. Hayter, M. Karababa, G. Logan and J. McFadden, "SGML-Based Markup for Literary Texts: Two Problems and Some Solutions", *Computers and the Humanities*, 22 [1988] 265–76). This problem received considerable attention within the TEI, the results of which are described in Barnard *et al.* in this volume.

[12] See Gaylord in this volume.

[13] See forx instance Johannson, Lavagnino and Mylonas, and Chisholm and Robey in this volume.

[14] See Cover and Robinson in this volume.

[15] See chapter 25 of TEI P3; see also Gaylord in this volume.

[16] See DeRose and Durand in this volume.

[17] See Dunlop, Ide and Véronis, and Langendoen and Simons in this volume.

[18] For meter, see Chisholm and Robey in this volume. 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.

[19] TEI P3 is available by anonymous ftp from ftp-tei.uic.edu (in pub/tei and its subdirectories), sgml1.ex.ac.uk (in tei/p3 and its subdirectories), TEI.IPC.Chiba-u.ac.jp (in /TEI/P3), and ftp.ifi.uio.no (in pub/SGML/TEI). Paper copies are available from offices in the U.S., England, and Japan. For more information contact C. M. Sperberg-McQueen at the University of Illinois at Chicago, Computer Center (M/C 135), 1940 W. Taylor St., Chicago IL 60612–7352, tei@uic.edu.

[20] The Japanese have formed a TEI committee to encourage TEI use in Japan and to coordinate Japanese input to the Guidelines.