The Association for Computers and the Humanities (ACH)

The Association for Computational Linguistics (ACL)

The Association for Literary and Linguistic Computing (ALLC)

Guidelines for Electronic Text Encoding and Interchange

Edited by

C. M. Sperberg-McQueen and Lou Burnard

TEI P2, Chapter 1
Text Encoding Initiative
Chicago, Oxford

March 9, 1993

Draft Version 2, March 9, 1993

Note

This is the current draft (as of March 9, 1993) of one section of the *Guidelines for Electronic Text Encoding and Interchange*, version 2 (also known as TEI P2), produced by the Text Encoding Initiative. This fascicle contains:

- a table of contents for TEI P2, indicating the status of each section as of the release date of this fascicle
- a draft for one section of TEI P2
- entries derived from this section for the reference list of tags of TEI P2 (if any)
- a User Response and Comment Form

Please feel free to send us your comments on the contents of this draft and their applicability to your needs. Comments may be sent to the TEI-L list server for general discussion, or directly to either of the TEI editors (adresses on the User Response and Comment form). Bear in mind in reading, however, that TEI P2 is a reference manual, not a tutorial in the use of electronic text, and unavoidably assumes either a certain amount of prior knowledge or patience with unfamiliar technical material. Note also that some matters which may appear obscure in one section may well be treated in more detail elsewhere in the full text of TEI P2.

TEI P2, like P1, is a working draft. The observant will find several typographic and presentational flaws in the current text: for example, cross-references are not properly resolved in the LaTeX versions and some parts of the reference material are not yet complete. We hope to rectify as many as possible of these flaws in the coming months, but have decided not to allow them to delay any further the release of the draft. We hope that those interested in the Guidelines will find the current state of the text readable and usable enough for the purpose of substantive comment. As far as possible, all comments received on this section will be taken into account in the preparation of the next version of the Guidelines, which is due by the end of 1992.

Updates on the status of the draft as a whole will be distributed automatically to all subscribers of the Listserv list TEI-L. To subscribe to TEI-L, send electronic mail to the address LISTSERV@UICVM (or Listserv@uicvm.uic.edu) containing the single line

```
subscribe tei-l J. Smith
```

(substituting your name for “J. Smith”). TEI-L is also the appropriate place to pose questions or offer public comments on the TEI guidelines and other relevant issues. The TEI-L file server contains all the sections of TEI P2 thus far released, as well as other TEI materials. Its contents are shadowed at various other sites around the world; for more information consult one of the following documents:

- TEI ED J8, “Obtaining the Second Version of the TEI Guidelines (TEI P2)” (describes how to retrieve electronic copies of TEI P2 and the various formats they are available in)

- TEI ED J9, “Obtaining Paper Copies of the Second Version of the TEI Guidelines (TEI P2)” (describes how to request paper copies of TEI P2, for those without electronic mail access)

These documents are available from the TEI-L file server or through the editors at the addresses given on the User Response and Comment Form.

This document is made possible in part by financial support from the U.S. National Endowment for the Humanities, an independent federal agency; Directorate General XIII of the Commission of the European Communities; the Andrew W. Mellon Foundation; and the Social Science and Humanities Research Council of Canada. Direct and indirect support has also been received from the University of Illinois at Chicago, the Oxford University Computing Service, the University of Arizona, the University of Oslo, Queen's University (Kingston, Ont.), and Ohio State University. We acknowledge assistance with software acquisition from the following firms: SEMA Group and Yard Software (for Mark-It and Write-It), Software Exoterica (for CheckMark and Xtran), SoftQuad, Inc., (for Author/Editor and RulesBuilder), Electronic Book Technology (for demonstration copies of DynaText), Xerox Corporation (for Ventura Publisher).

Overall Table of Contents for TEI P2

The following list shows the expected contents of TEI P2 as a whole and indicates the current status of each chapter.

Contents

Part 1

Introduction

Chapter 1
About these Guidelines

These Guidelines have been developed by the Text Encoding Initiative (TEI); see 1.5, "Historical Background," on page 15. They are addressed to anyone who works with any text in electronic form. They provide a means of encoding those features of a text which need to be identified in some way in order to aid the processing of that text by computer programs. Such encoding is often called "mark-up" or "tagging" and the term *encoding scheme* or *markup language* denotes the rules which govern the use of markup in a set of encodings.

The TEI Guidelines are intended to codify the form, content, and interpretation of textual features for a wide variety of purposes. They are intended for use in interchange between individuals and research groups using different programs and computer systems over a broad range of applications. They can also be used for the local storage of text which is to be processed with multiple software packages with different input formats. And since they contain an inventory of the features most often found useful for text processing, the Guidelines also provide help to those creating electronic texts.

It should be noted that the present document is a technical reference manual for a markup language designed to solve a large number of occasionally complex problems; it is not a tutorial, and should not be read as one. No user of the scheme described here is likely to need to be familiar with the whole of it. Separate tutorials on SGML and the TEI are being developed, which will include both introductory background material for novice readers and more detailed information appropriate to particular application areas.

This introductory chapter begins by describing informally the ideas underlying our approach to the encoding of textual material in electronic form, (section 1.1, "Texts and Their Electronic Representation,"), and specifying the kinds of usage envisaged for the present recommendations (section 1.2, "Intended Applications," on page 8). This is followed by a brief summary of its contents (section 1.3, "Structure of This Document," on page 11), and a definition of the stylistic conventions and notation used throughout the body of the text (section 1.4, "Notational Conventions Used in This Draft," on page 13). Finally, section 1.5, "Historical Background," on page 15 gives a brief account of the process by which the present draft came into being, and its likely development in the future.

1.1 Texts and Their Electronic Representation

This section briefly summarizes some of the design principles and objectives of the work which has lead to the present publication.

1.1.1 What is 'Text'?

By *text* we mean any kind of written or spoken language. Our primary focus is on the raw material of literary, linguistic, and other textual scholarship of all kinds, in any natural language, of any date, in any literary genre or text type, without restriction on form or content. We treat both continuous materials ("running text") and discontinuous materials such as dictionaries and linguistic corpora.

The interests of those who work with electronic text themselves cover a broad and varied field. Included are linguists of all kinds, literary scholars, philosophers, historians, anthropologists, lexicographers, psychologists, discourse analysts and computer scientists. The TEI Guidelines should be equally useful to scholars in any of these disciplines and to librarians who maintain and document electronic materials, no matter how large or small the text they are working with. Though principally directed to the needs of the scholarly research community, the Guidelines are not restricted to esoteric academic applications. Publishers and others entering the market for electronic texts may also find that many of their needs are met here.

1.1.2 Views of Text

The TEI Guidelines are built on the assumption that there is a common core of textual features shared by virtually all texts and virtually all serious work on texts. Nevertheless, views of text vary widely. In different contexts, texts may be considered as sequences or hierarchic arrangements of:

- 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 (characters in specific fonts, laid out and justified in a particular style)
- linguistic objects (graphemes or phonemes, or at a higher level series of morphemes or lexical items or phrases or sentences)
- literary objects (stanzas, cantos, acts, chapters, sections, etc.)
- rhetorical objects (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 layers of interpretation, re-interpretation, and commentary)

In many situations more than one view of a text is needed, for example the physical and the linguistic or the formal and the rhetorical. Consequently, no absolute recommendation to embody one specific view of text can apply to all texts and all approaches to it. Unlike some existing encoding schemes, these Guidelines define a general-purpose encoding scheme which attempts to make it possible to encode multiple views of text.

The attempt to accommodate all possible characteristics of text within the same general-purpose encoding scheme has two major advantages. First, general-purpose tagging can help reduce the costs of creating electronic texts, since the same text can be used for many purposes. Indeed some texts are now produced with the explicit goal of creating generally useful tools for scholarship, rather than serving one particular research project.

Second, a general-purpose scheme can help ensure the intellectual freedom of the researchers encoding a text. Any encoding scheme gives names to, and provides methods of recording the existence of, some specific set of textual features. Encoding schemes specialized for one particular view of text will reflect the particular bias of that view, and a general-purpose scheme, by contrast, provides methods for encoding as broad a variety of textual features as the consensus of the community permits, and also allows the encoder of the text to extend the markup language by adding new textual features to its vocabulary.

The TEI Guidelines give names for particular features, but with relatively rare exceptions they place no restrictions on the encoder as to what features are important and what are unimportant. The task of deciding which of the features defined here are to be encoded falls to those responsible for the actual encoding of a specific text.

In some cases these Guidelines provide multiple possible encodings for what appear to be identical textual phenomena. A typical case is that of a word, rendered in italics. One encoder may wish to record simply that the passage is rendered typographically using an italic rather than a roman font. Another encoder may want to register rhetorical or linguistic distinctions and identify it as a title, a technical term, or an emphatic word. Either view, or both, may be encoded using the scheme presented here.

1.1.3 Problems of Encoding Text

Any general-purpose scheme for encoding texts must address several basic technical problems in the representation of texts in electronic form:

1. how to represent the individual characters of a text, especially those not included in widely implemented standard character sets, such as
    a. the accents needed for French, German, Spanish and many other languages written in Latin script
    b. the special consonants and vowels needed for other languages such as Polish, Turkish, or the medieval forms of some languages (e.g. Old and Middle English)
    c. the alphabetic symbols used in writing systems other than the Latin script
    d. punctuation marks, both those peculiar to some writing systems, such as the Greek colon, and those which are either not distinguished or not available in modern data processing systems, such as opening and closing quotation marks, differing forms of dashes or hyphens, distinct forms of abbreviations in manuscripts and so forth
    e. non-alphabetic symbols such as mathematical symbols, the signs of the Zodiac, chemical and astrological symbols, etc.
    f. the symbols or ideograms used in non-alphabetic writing systems, phonemic transcriptions etc.
2. how to record the basic logical structure of the source text (e.g. its division into book, chapter, and paragraph, or into play, act, scene, and line), so that passages found by a search can be located in a printed copy of the text

3. how to reduce text with footnotes, apparatus, parallel printing, etc. to the single linear sequence required by most computer file systems
4. how to record features which are explicitly marked in the text such as titles, authors, dates, and sentence boundaries
5. how to record features not so explicitly marked, such as scansion of verse, metaphors and allusions
6. how to record the results of analyzing and interpreting the features and content of the text according to particular research goals
7. how to distinguish comments and other extraneous matter in the electronic text from the transcription itself

The TEI Guidelines provide specific names for many of the features mentioned above but (with relatively rare exceptions) make no suggestions or restrictions as to their relative importance. It is the encoder's responsibility to decide what to encode. The philosophy of the Guidelines is “if you want to encode this feature, do it this way”: but very few features are mandatory. Some texts may therefore have very little encoding; others which are more complex may require many more encoding tags. We see the encoding process as incremental. Tags for different features may be added to texts as new scholars work on those texts for different applications.

These Guidelines use rules and techniques which are expressed in the Standard Generalized Markup Language (SGML), which is described in more detail in chapter 2, “Concise Summary of SGML,” [in separate fascicle]. Particular features of the notation used are described below in section 1.4, “Notational Conventions Used in This Draft,” on page 13.

SGML however only provides the syntax in which we put forward our solutions to the problems outlined above. The specific tags and rules governing their use which are described in the TEI Guidelines necessarily reflect the research orientation of the community which together developed them. Hence the careful attention paid to documenting the copy text and the process of transcription (see chapter 5, “The TEI Header,” [in separate fascicle]), and also the detailed suggestions for some areas perhaps less well-covered by other encoding schemes, including many which also use SGML. Examples include text types other than conventional prose (see part III), text-critical annotation (chapter 21, “Text Criticism and Apparatus,” [in separate fascicle]), and general-purpose methods for recording analysis and interpretation of a text (chapters 16, “Segmentation and Alignment,” [in separate fascicle], 17, “Simple Analytic Mechanisms,” [in separate fascicle] and 18, “Feature Structure Analysis,” [in separate fascicle]).

1.1.4 Encoding and Interpretation

In one respect, these Guidelines take the possibly controversial view that a researcher's analysis or interpretation of the text may be reflected in the electronic encoding of that text. It is sometimes suggested that encoders of electronic text avoid bias toward particular theoretical positions by limiting themselves to “objective” transcription and eliminating all subjective or interpretive material from the encoding. The TEI Guidelines reflect a different view. The distinction between objective and subjective information is often blurred and sometimes itself the subject of scholarly research. The syntax of SGML ensures that some encodings can be ignored for some purposes. We also provide a means of documenting the interpretive encoding in such a way that a user of the text can know the reasoning behind that interpretation. The accuracy and reliability of that reading is for the individual user of the text to determine.

1.1.5 Fidelity to and Recoverability of the Source

Any encoding constitutes a representation, in electronic form, of the text encoded. Whatever work is to be done with the encoding, the encoding must be faithful, in some useful sense, to the text being represented. This requirement may be expressed in the dictum, occasionally proposed as a rule for encoding, that the original text must be recoverable from the encoding.

In general, however, representations of any object do not capture all aspects of the object represented; they simplify the object by capturing only its salient features and omitting negligible features. The creator of an electronic version of a text, therefore, must actively choose the aspects of the original to record depending on the objectives to be satisfied. Here we can offer some guidance with examples taken from printed texts.

One might, for example, provide "image fidelity" by recording the physical characteristics of the text in great detail, noting each page and line break, each change of type, the specific fonts used (point size, leading, family, style, maker, etc.), and the position and text of each header and footer line on the page. Carried to the logical extreme, such a fidelity to the physical appearance of the text would allow the creation of a facsimile edition from the electronic form of the text. Less extremely, one could provide "typographical fidelity" by registering the distinctions typically made by a copy editor's instructions to the typesetter in conventional publishing, like information about font shifts and layout.

"Rhetorical fidelity" would record the linguistic or rhetorical phenomena signaled by typographic conventions. Details of the typography might be associated systematically with the underlying linguistic or rhetorical phenomena such as emphasis, sentence structure, etc. (or might in some cases be discarded).

It is also possible to abandon "fidelity" to the printed source, and focus entirely on the information content of the text being encoded. In the case of dictionaries or other reference works, a user might move away from the representation of the printed reference work, and toward information in the form of a database completely re-ordered for convenience of processing with inconsistencies in the source text eliminated, etc. Such an encoding might retain some information about the printed form, of course, if only to ensure that the printed form of the source reference work was recoverable at some level.

The TEI Guidelines do not specify a particular approach to the problem of fidelity to the source text and recoverability of the original; such a choice is the responsibility of the text encoder. It is strongly recommended that the TEI header be used to give an account of which aspects of the original text are recoverable from the electronic encoding and which are not recoverable; the TEI header is described in chapter 5, "The TEI Header," [in separate fascicle]. The current version of these Guidelines, however, provides a more fully elaborated set of tags for markup of rhetorical, linguistic, and simple typographic characteristics of the text than for detailed markup of page layout or for fine distinctions among type fonts or manuscript hands.

1.2 Intended Applications

We envisage three primary functions for these Guidelines

- guidance for individual or local practice in text creation and data capture.
- support of data interchange
- support of application-independent local processing

These three functions are so thoroughly interwoven in practice that it is hardly possible to address any one without addressing the others. However, the distinction provides a useful framework for discussing the possible role of the Guidelines in work with electronic texts.

1.2.1 Use in Text Creation

The description of textual features found in the chapters which follow should provide a useful checklist from which scholars planning to create electronic texts should select the subset of features suitable for their project.

In some cases we have found that the consensus of the text-processing community allows us to make specific recommendations about preferred practices. Where a given feature is generally found useful, for example, the tag for that feature is recommended for general use; where it is found not worth tagging, that fact is also noted. In other cases we have been able to establish more general methodological principles. For still others, the tagging or omission of other features is left to the discretion of the individual working with the text.

For those new to the field, such recommendations are of obvious use. For those who have very specific ends in view and who already know what features must be captured to serve those ends, the recommendations may show how the text may be made useful for other purposes. The recommendations, while seriously intended, should be distinguished rigorously from the requirements of the TEI encoding scheme, which are comparatively few.

General problems of text interpretation and systematic analysis of texts have not been treated in this document, beyond the cursory treatment given them in this chapter. Similarly problems specific to text creation or text “capture” have not been considered explicitly in the pages which follow. For purposes of the TEI interchange format and for use of SGML, it does not matter how a text is created or captured: it can be typed by hand, scanned from a printed book or typescript, read from a typesetter's tape, or acquired from another researcher who may have used another markup scheme (or no explicit markup at all).

We include here only some general points which are often raised about SGML and the process of data capture.

To begin with, SGML can appear distressingly verbose, particularly when (as in these Guidelines) the names of tags and attributes are chosen for clarity and not for brevity. Editor macros and keyboard shorthands can allow a typist to enter frequently used tags with single keystrokes. Special-purpose software may be purchased which scans word-processor or scanner data and inserts SGML tags. And the techniques described in chapter 32, “Modifying TEI DTDs,” [in separate fascicle] may be used to give shorter names to the tags being used most often. It should also be noted that the examples in this text are chosen to exhibit the markup as compactly as possible, and thus have denser markup than will be typical in many texts.

The SGML standard provides ways of abbreviating, omitting, or otherwise *minimizing* the amount of markup which need be explicitly provided in a text. They are all forbidden in the TEI interchange format because their use complicates processing; this does not however preclude their use in local processing, where this is felt appropriate or desirable.

1.2.2 Use for Interchange

When the TEI Guidelines are used for interchange, it is expected that researchers using other encoding schemes in their work will translate outgoing data from such schemes into the scheme described by these Guidelines, and similarly translate incoming data from the scheme described here into those used internally. For such translations to be carried out without loss of information, the scheme proposed here must be as expressive (in a formal sense) as any encoding scheme now known to be in wide use for textual research. To ensure that this is the case, a set of extension techniques is provided (see chapter 32, “Modifying TEI DTDs,” [in separate fascicle]) which makes possible the addition of extra tags, the renaming of existing tags and certain kinds of redefinition. However the intention has been to minimize any need for recourse to such extensions. To translate between any pair of encoding schemes implies

1. identifying the sets of textual features distinguished by the two schemes
2. determining where the two sets of features correspond
3. creating a suitable set of mappings

For example, to translate from encoding scheme X into the TEI

1. Make a list of all the textual features distinguish in X.
2. Identify the corresponding feature in the TEI scheme. There are three possibilities for each feature:
   a. feature exists in both X and the TEI
   b. X has a feature which is absent from the TEI
   c. X has a feature which corresponds with more than one feature in the TEI scheme

   The first case is unproblematic. The second requires an extension to the TEI scheme, as described in chapter 32, “Modifying TEI DTDs,” [in separate fascicle]. The third requires that a consistent choice be made. The algorithm used to make that choice should be documented in the TEI header.

3. Using the table of equivalences so generated, a simple translation can be carried out between X and the TEI.

Translating from the TEI into scheme X follows the same pattern, except that if a TEI feature has no equivalent in X, and X cannot be extended, information must be lost on translation.

Similar procedures may be followed where the TEI scheme is to be used as an interlanguage for interchange among several different sites or applications, although the degree of TEI-conformance may vary.

In the simplest case, where two sites or individuals exchanging texts know each other and know or can inquire what equipment the other is using, these Guidelines serve primarily as documentation for a file format, which can be referred to without actually being transmitted together with the file. In the general case, where sender and recipient cannot communicate such information, a stricter degree of *TEI conformance* will be required for loss-free interchange.

The rules defining such strict conformance to the Guidelines are given in some detail in chapter 31, "TEI Conformance," [in separate fascicle]. The *interchange format* defined there requires that an electronic text:

1. adhere to the SGML declaration and the SGML document type declarations reproduced in the appendix, unless modified or extended as described in chapter 32, "Modifying TEI DTDs," [in separate fascicle]. (The concepts of the SGML declaration and the SGML document type declaration are explained in section 2, "Concise Summary of SGML," [in separate fascicle]. The provisions of the SGML declaration to be used for TEI files are discussed in section 2, "Concise Summary of SGML," [in separate fascicle].)
2. provide external documentation as described in chapter 30, "Tag Set Declaration," [in separate fascicle] for all elements not defined in the TEI Guidelines, specifying a formal name (generic identifier) and a corresponding full natural-language name, describing its meaning and usage, specifying its legal content and also any attributes it may use.
3. adhere to the requirements of the TEI header in providing bibliographic identification of the text and description of the encoding practices used (as described in chapter 5, "The TEI Header," [in separate fascicle]).

Note that the interchange format makes no formal restriction on the character set to be used in interchange, as this will depend on the medium of interchange and the local character sets in use by sender and receiver. For further information, refer to chapter 34, "Use of TEI Encoding Scheme in Interchange," [in separate fascicle].

1.2.3 Use for Local Processing

Machine-readable text can be manipulated in many ways; our aim has been to avoid assuming too much about what the reader of these Guidelines will do with texts marked up according to these rules. Some possible activities include:

- edit texts (e.g. word processors, syntax-directed editors)
- edit, display, and link texts in hypertext systems
- format and print texts using desktop publishing systems, or batch-oriented formatting programs
- load texts into free-text retrieval databases or conventional databases
- unload texts from databases as search results or for export to other software
- search texts for words or phrases
- perform content analysis on texts
- collate texts for critical editions
- scan texts for automatic indexing or similar purposes
- parse texts linguistically
- analyze texts stylistically
- scan verse texts metrically
- link words of a text to images of the objects named by the words (as in a hypertext language-teaching system)

These applications cover a wide range of likely uses but are by no means exhaustive. The aim has been to make the TEI Guidelines useful for encoding the same texts for different purposes. We have avoided anything which would restrict the use of the text for other applications. We have also tried not to omit anything essential to any single application.

Use of the TEI scheme for local processing involves the acquisition of software able to deal with documents in TEI format. Since the TEI interchange format uses only a rather small subset of SGML (a formal grammar for which appears in chapter 42, “Formal Grammar for the TEI-Interchange-Format Subset of SGML,” [in separate fascicle]), many programmers may find it simple to develop *ad hoc* software to deal with TEI texts. Others will find it simpler to use existing text-manipulation software to work with TEI texts, or to acquire general-purpose SGML software for validating and processing documents. The increasingly wide acceptance of SGML as a document-processing standard means that such software will continue to become more and more widely available.

1.3 Structure of This Document

This document is the (draft) reference manual for the TEI encoding scheme. It will be complemented by a series of tutorials in text encoding (document TEI U1 et seq.), and a case book of extended examples with discussion of the rationale for various markup choices (TEI T1). Readers seeking an introduction to text markup and the use of the TEI encoding scheme in a specific area should consult an appropriate tutorial; those already familiar with the scheme and interested in seeing examples of its application should consult the case book.

This document is intended to provide an authoritative statement of the requirements and usage of the TEI encoding scheme. It includes numerous small examples, but few readers will find it convenient to learn the encoding scheme by reading this document alone.

Part I provides some relevant background information about the Guidelines themselves (in this chapter); a brief technical review of SGML, the Standard Generalized Markup Language (chapter 2, "Concise Summary of SGML," [in separate fascicle]); and a description of how the TEI document type declarations are organized (chapter 3, "Structure of the TEI Document Type Declarations," [in separate fascicle]).

Part II provides a systematic treatment of issues common to all text types: character representation (chapter 4, "Characters and Character Sets," [in separate fascicle]); in-file documentation of the text (chapter 5, "The TEI Header," [in separate fascicle]); tags for text features found in all sorts of text: lists, notes, emphasis, quotations, cross-references, technical terms, names, dates, numbers, etc. (chapter 6, "Tags Available in All TEI DTDs," [in separate fascicle]); and a definition for the default structure of all TEI documents (chapter 7, "Default Text Structure," [in separate fascicle]).

Part III documents various *base tag sets*: these include specialized tags for prose, for verse, for drama and other performance materials, for spoken materials, as well as for letters and memoranda, printed dictionaries, and terminological data. Additional sections discuss user-defined and mixed base tag sets. A TEI DTD must use one and only one base tag set, unless one of the "mixed" bases is used.

Part IV documents various *additional tag sets*, which may be included or excluded, as appropriate. Topics covered include a variety of approaches to the analysis and interpretation of texts, and include representations for hypertextual links and other non-hierarchic structures, as well as specialized tags for the encoding of physical description, critical editions, and language corpora.

Part V defines certain specialized *auxiliary document types*, used to encode information about the way that texts have been encoded, specifically: the TEI header regarded as a distinct document; the TEI Writing System Declaration; the Feature System declaration; and the Tag Set Documentation.

Part VI contains a number of technical discussions of a more specialist interest. Topics covered include the notion of formal *conformance* to the TEI Guidelines; the controlled user-modification of the TEI DTDs; practical aspects of the use of TEI markup both in local processing and in interchange; and the relationship of TEI markup to other markup standards.

Part VII consists of an alphabetical reference list of all elements and element classes defined in the TEI encoding scheme.

Part VIII provides further reference material: specifically, the full TEI document type declarations, a set of standard Writing System Declarations, a sample Feature System Declaration for basic grammatical annotation, sample tag documentation, and a formal grammar for the subset of SGML used in the TEI interchange format.

In the back matter, a bibliography lists all works cited in the text of the Guidelines. A User Response and Comment form serves as a reminder that the document in hand is a draft, for which public reaction is vigorously solicited.

1.4 Notational Conventions Used in This Draft

This section describes the typographic and stylistic conventions used throughout this document.

SGML elements mentioned in the text take the form **<name>**, where "name" is the *generic identifier* of the element. Sample tags mentioned in the text are displayed in the form **<name att=value att2='value two'>**. Where the elements thus mentioned are part of the TEI encoding scheme, they are included in the Temporary Index.

These Guidelines distinguish encoding practices, and SGML elements, which are required, recommended, or optional. The phrases "must", "is required to", etc., mark practices and tags which are required for TEI conformance. The phrases "should", "it is recommended that", "it is preferable to ...", etc., are used in describing practices which are recommended but not required for TEI conformance. Modal verbs like "may", "might", etc., mark practices which are strictly optional. Qualifying phrases like "if desired", "where appropriate", or "under some circumstances" are used when some tag or practice described may be desirable or acceptable under some circumstances and not under others.

In the reference section in Part VII, elements and their attributes are all classed as one of:

required: unconditionally required in a TEI-conformant document
mandatory when applicable: required under the appropriate conditions; may be omitted if not applicable
recommended: recommended unless there are good reasons, in the given circumstances, against it
recommended when applicable: recommended under some circumstances (which should be clear from context)
optional: strictly optional

This reference section includes cross-references to the chapter or section of the main text within which each element is discussed. Most sections of the main text in which elements are defined begin with a descriptive list of the elements concerned in the following format:

<tag> : short description of the element marked by **<tag>**. Where appropriate this is followed by a list of significant non-global attributes for the element as follows:
- *attribute* : description of the attribute's meaning or usage, optionally followed by a list of suggested or legal values:
  - **value1**: meaning of value1
  - **value2**: meaning of value2

Not all attributes are always included in these lists; those which are shared with other elements in a class are usually listed separately, and those of relatively specialized interest are usually listed only in the reference section. The values of the attribute are introduced with one of the following formulaic phrases:

"Legal values include:" : The attribute cannot take values other than those given. Other values will cause SGML parsing errors.

"Suggested values include:" : The values listed constitute a set which should suffice for most purposes, and they should be used where appropriate. Developers of TEI-aware software should ensure that their software can process these values appropriately. In some cases, however, it is conceivable that other values might be necessary, so the SGML declaration for the attribute does not restrict legal values to those given. TEI-aware software should have reasonable fallback processing for values not in the list.

"Sample values include:" : The attribute can take any value; those listed are provided simply as examples of the kind of value possible.

Each list of elements is followed by some discussion of its semantics and usage, followed by one or more examples, taken wherever possible from real texts, and presented in the following format:

```
<p>A fly built a castle, a tall and mighty castle. ... </p>
```

All the examples are (or should be) legal SGML, but, because they are fragmentary, may not be parseable by SGML parsers without the required context. They also frequently make liberal use of white space to exhibit the logical structure of the SGML coding more clearly. Although this does not affect the SGML conformance of the examples, some users will prefer not to follow it in practice, since not all processors will ignore the extra white space. Examples may:

- show full start- and end-tags for all elements
- use empty end-tags (of the form **</>**) to close the most recently opened element
- omit end-tags (never start-tags) where they may legally be omitted

Attribute values are given indifferently in single quotes or double quotes; unquoted attribute values are sometimes used where SGML requires no quotation marks.

The examples thus demonstrate a variety of tagging styles, mostly aimed at making the tagging legible while also showing fairly explicitly where all elements begin and end. No claim is made or implied as to the appropriateness of the style adopted here for other purposes; in particular, those using SGML for local processing may often prefer to use empty end-tags or to omit tags more frequently than is shown in the examples.

After the examples and usage notes, the section typically concludes with a DTD fragment containing the formal declarations for the elements described. Each DTD fragment is given a heading, and may contain element and attribute list declarations, entity declarations, parameter entity references, comments, and references to DTD fragments in other sections. The DTD fragments of a single chapter almost invariably belong to the same DTD file, the structure of which is typically described (with references to the included fragments) in one of the first or last sections of the chapter.

The DTD fragments are identical to the DTDs distributed with these Guidelines, with the following exceptions:

- in the text, the DTD fragments appear in the order dictated by organization of this document; the actual DTD files may re-order the material slightly. This is indicated in the text by references from one DTD fragment to another.
- the DTD fragments in the text show the generic identifiers of all elements using the standard English names assigned in this document; the actual DTD files use parameter entities for all generic identifiers, so that elements can be conveniently renamed, as described in chapter 32, “Modifying TEI DTDs,” [in separate fascicle].
- the actual DTD files include conditional marked sections surrounding the element and attribute list declaration for each element, to ensure that elements can conveniently be suppressed or redefined, as described in chapter 32, “Modifying TEI DTDs,” [in separate fascicle]. The fragments in the text suppress the marked-section-open and marked-section-close markup.

What appears in the text, therefore, as:

```
<!ELEMENT blort         - O  (#PCDATA | grackle | granfalloon)* >
<!ATTLIST blort              %a.global;
          type               CDATA              #IMPLIED
          resp               CDATA              #IMPLIED       >
```

will appear thus in the actual DTD file:

```
<!ENTITY % blort 'INCLUDE' >
<![ %blort [
<!ELEMENT %n.blort;     - O  (#PCDATA | %n.grackle |
                             %n.granfalloon)*                  >
<!ATTLIST %n.blort;          %a.global;
          type               CDATA              #IMPLIED
          resp               CDATA              #IMPLIED       >
]]>
```

For further discussion, see chapter 3, “Structure of the TEI Document Type Declarations,” [in separate fascicle], or chapter 32, “Modifying TEI DTDs,” [in separate fascicle].

1.5 Historical Background

These Guidelines are the result of several years' development by the an international co-operative research project known as the Text Encoding Initiative (TEI). This section outlines the progress of that project and its likely future development.

1.5.1 Origin and Development of the TEI

The Text Encoding Initiative grew out of a planning conference sponsored by the Association for Computers and the Humanities (ACH) and funded by the U.S. National Endowment for the Humanities (NEH), which was held at Vassar College in November 1987. At this conference some thirty representatives of text archives, scholarly societies, and research projects met to discuss the feasibility of a standard encoding scheme and to make recommendations for its scope, structure, content, and drafting. During the conference, the Association for Computational Linguistics and the Association for Literary and Linguistic Computing agreed to join ACH as sponsors of a project to develop the Guidelines. The outcome of the conference was a set of principles, listed in the following section, which determined the further course of the project.

The Text Encoding Initiative proper began in June 1988 with funding from the NEH, soon followed by further funding from the Commission of the European Communities, the Andrew W. Mellon Foundation, and the Social Science and Humanities Research Council of Canada. Four working committees, composed of distinguished scholars and researchers from both Europe and North America, were named to deal with problems of text documentation (resulting largely in chapter 5, "The TEI Header," [in separate fascicle]), text representation, text analysis and interpretation (together responsible for most of what has become parts II, III, and IV), and metalanguage and syntax issues (largely responsible for part VI).

A first draft version (1.0) of the Guidelines was distributed in July 1990 under the title *Guidelines for the Encoding and Interchange of Machine-Readable Texts*, with the TEI document number TEI P1. With minor changes and corrections, this version was reprinted as version 1.1 in November 1990.

Extensive public comment and further work on areas not covered in version 1 resulted in the drafting of the current version, TEI P2, distribution of which began in April 1992. This version includes substantial amounts of new material, resulting from work carried out by several specialist working groups, set up in 1990 and 1991 to propose extensions and revisions to the text of P1. The overall organization, both of the draft itself and of the scheme it describes, has been entirely revised and reorganized in response to public comment on the first draft, which this version supersedes.

1.5.2 Design Principles of the TEI

The planning conference mentioned in section 1.5.1, "Origin and Development of the TEI," agreed on the following statement of principles:

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.

These basic principles are expounded in various documents of the Text Encoding Initiative (notably TEI EDP1 and TEI EDP2) and the interested reader is directed to those documents for further discussion.

The mandate of creating a common interchange format requires the specification of a specific markup syntax as well as the definition of a large predefined tag set and the provision of mechanisms for extending the markup scheme. These mandates are fulfilled by the provision of full SGML document type declarations for the scheme described here and by the extension mechanisms described in chapter 32, “Modifying TEI DTDs,” [in separate fascicle]. The mandate to provide guidance for new text encodings (“suggest principles for text encoding”) requires that recommendations be made as to what textual features should be recorded in various situations. This mandate is fulfilled by the explicit specification, in the reference section for each tag, that the tag is *required*, *mandatory when applicable* but otherwise omissible, *recommended* generally, *recommended when applicable* but not always applicable, or *optional*. More generally, the principles of text encoding recommended by these Guidelines are described in this chapter and embodied in the encoding scheme presented in the chapters which follow.

In designing the tag set and formulating the recommendations, the following design goals have been paramount. These Guidelines are intended to:

1. suffice to represent the textual features needed for research
2. be simple, clear, and concrete

3. be easy for researchers to use without special-purpose software
4. allow the rigorous definition and efficient processing of texts
5. provide for user-defined extensions
6. conform to existing and emergent standards

The TEI Guidelines do not completely fulfill the first design goal: there will always be areas of scholarly research not yet explicitly addressed, while those that are discussed here have (in many cases) yet to be put to the test. Further revisions and extensions of the Guidelines are therefore to be expected in the light of experience; see further 1.5.3, “Future Developments.”

The simplicity and ease of use of the Guidelines are best left to the reader to judge; examples throughout the text and appendices should, however, make clear that the markup here described does allow the rigorous definition and processing of textual objects and can be used without special software -- although the experience of design has suggested more strongly than before how useful it is to have software capable of exploiting SGML markup.

The rules and recommendations made in this document do conform to the salient international standards (notably ISO 8879, which defines the Standard Generalized Markup Language, and ISO 646, which defines a standard seven-bit character set in terms of which the recommendations on character-level interchange are formulated.

1.5.3 Future Developments

This is version P2 of the TEI Guidelines for Text Encoding and Interchange. It will be revised based upon public comment on this draft and further work by the working groups responsible for drafting it. The resulting third draft (which will have the TEI document number TEI P3) will be presented to the Advisory Board of the Text Encoding Initiative for approval and endorsement in June 1993.

After revision in accordance with suggestions of the Advisory Board, the third version of the Guidelines will be officially published.

Work on areas still not satisfactorily covered in TEI P3 will continue, and resulting recommendations will be issued as supplements to the published Guidelines. Work is expected to continue in at least the following areas:

- linguistic description and grammatical annotation
- historical analysis and interpretation
- base tag sets for further document types
- manuscript analysis and physical description of the copy text

The encoding recommended by this draft may be used without fear that future versions of the TEI scheme will be inconsistent with it in fundamental ways. The TEI will be sensitive, in revising this draft, to the possible problems which revision might pose for those who are already using this draft. Wherever consistent with the long-term goals of the project, consistency with this draft will be preserved in future revisions.

Usability Report and Comment

If you have comments of any kind on the current draft of the TEI Guidelines, please complete this form and return it to one of the following addresses:

C. M. Sperberg-McQueen, Computer Center (M/C 135), University of Illinois at Chicago, 1940 W. Taylor Room 124, Chicago IL 60612-7352, USA. Bitnet: U35395@UICVM, Internet: U35395@uicvm.uic.edu

Lou Burnard, Oxford University Computing Service, 13 Banbury Road, Oxford OX2 6NN, UK. Fax: +44 (865) 273275, Janet: LOU@UK.AC.OXFORD.VAX Internet: LOU@vax.oxford.ac.uk

Further copies of this form may be downloaded from LISTSERV @ UICVM (or listserv @ uicvm.uic.edu). Send a mail message containing the line 'GET TEIURC MEMO TEI-L'

Your name:
Your postal address:

Your e-mail address (if any):
Your occupation:
Your academic background:
Your immediate reactions to the Guidelines (please tick or cross one only):

Relevance to your present concerns or interests: high medium low
Importance to your future research plans: high medium low
Comprehensibility/usability of current draft: high medium low

Detailed technical comment: [Please feel free to comment on any part of the Guidelines or in general. Comments relating to specific parts of the text are most helpful if they include a section or chapter number]

Thank you!

Temporary Index