28 Conformance

The recommendations in this chapter are likely to be substantially revised at the next release.

The notion of *TEI conformance* is intended as an aid in describing the format and contents of a particular document or set of documents. These concepts are expected to be useful in:

- agreements for the interchange of documents among researchers
- agreements for the deposit of texts in archives and their distribution from archives
- describing the documents to be produced by or for a given project
- defining the classes of documents accepted or rejected by a given piece of software

This chapter describes the areas in which these terms are defined and specifies their meaning. It also proposes other terms for related concepts and points out some dangers in the careless use or application of these terms.

28.1 Definitions of Terms

The terms described here should be considered technical terms for users and implementors of the TEI Guidelines and should be used only in the senses given and with the usages described.

28.1.1 TEI-Conformant Document

A document is *TEI-conformant* if it is either in *TEI local processing format* or in *TEI interchange format*. A full description of the document should specify which format it is in.

The term *TEI conformance* does not apply to software: programs can be usefully described as *accepting* or *validating* TEI-conformant documents or some subset of TEI-conformant documents, but the TEI defines no required processing model against which software could be measured. Programs are thus not themselves conformant or non-conformant and should not be so described.

28.1.2 TEI Local Processing Format

A document is in *TEI local processing format* if

- it is a valid XML document;
- or, alternatively, it is a conforming SGML document with an appropriate legal SGML declaration;
- it uses the document type declarations provided by the TEI, either without modifications or with all modifications effected by inclusion in the DTD subset as described in section 28.3 *Modifications to TEI Document Type Declarations*;
- all modifications to meaning or use of defined tags, and all new tags, are documented in TEI Tag Set Declarations which accompany the document, as defined in chapter 27 *Tag Set Documentation*;
- it includes, in the TEI header, all the elements required by the TEI declarations for the TEI header;
- it contains no non-SGML, non-XML markup other than declared notations for graphics, tables, figures, etc. That is, unless a declared notation is in use, the semantics of any content character in the document are exhausted by its identity as a graphic character.

A TEI-local-processing-format document may be described as requiring *DTD extensions* if it modifies the TEI-supplied DTDs (or in the case of SGML, the SGML prolog) in any of the ways described under 28.3 *Modifications to TEI Document Type Declarations*.

The following terms are synonymous: *document in TEI local processing format*, *TEI local-processing document*, and *TEI local-processing-conformant document*.

28.1.3 TEI Interchange Format

A document is in *TEI interchange format* if it conforms to the TEI local-processing format and if further:

- it is valid XML; or is SGML using either the predefined SGML declaration for TEI interchange documents or one which differs from it only in ways allowed by section 28.2 *Modifications to TEI SGML Declaration*;
- it makes no use of any of the following SGML constructs:[170]
  - short references
  - the RANK feature
  - omission of generic identifiers in start- and end-tags[171]
  - keywords other than INCLUDE, IGNORE, and CDATA on a marked section
- it includes no SUBDOC subordinate document by means of an entity reference embedded directly within content data (SUBDOCs must be included by giving the entity reference as the value of an attribute);
- it does not provide different definitions for the same entity in different document types.

A TEI-interchange-format document may be described as requiring *DTD extensions* if its DTD is modified in any of the ways described in section 28.3 *Modifications to TEI Document Type Declarations*.

The following terms are synonymous: *document in TEI interchange format*, *TEI interchange document*, and *TEI interchange-conformant document*.

28.1.4 TEI Packed Interchange format

A document is in TEI packed interchange format with a given *transmission character set* and a given *transmission entity set* if all of the following are true:

- all separate entities in the document are packed into a single entity (file) in a manner conforming to ISO 9069 (SDIF) or to some other TEI-authorized form;
- all characters occurring in SGML or XML names (generic identifiers and attribute names) occur within the transmission character set;
- all characters in the document content and attribute values either occur within the transmission character set or are represented by an appropriate entity reference using an entity name included in the transmission entity set;
- the transmission character and entity sets are named in the header of the packed file and in any accompanying paper documentation.

Any pre-transmission processing required to convert a document to meet the above requirements for conformance to the TEI-interchange-format is called *packing*.

With prior agreement between parties to an exchange, interchange documents may use character code set switching as defined in ISO 2022, its national analogues, or successor standards.

A full description of a document in TEI packed interchange format *must* specify the transmission character set and the transmission entity set used in the document.

28.1.5 TEI Recommended Practice

A document follows *TEI recommended practice* if:

- it is a TEI-conformant document;
- wherever the guidelines say to prefer one encoding practice to another, the preferred practice is used;
- all textual features which the guidelines recommend be captured are in fact encoded.

[170] The definition of interchange format may be changed to eliminate the few remaining SGML features that are not in XML (primarily attribute minimization).

[171] This is one of several abbreviations allowed by the SHORTTAG feature; the others (omission of attribute names under certain circumstances and omission of non-required attribute values) are allowed by the current release of the Guidelines, but users are cautioned that this may be changed at a subsequent release, in the interests of XML-conformance.

28.1.6 TEI Abstract Model

A document follows the *TEI abstract model* if it tags the features specified in the TEI documentation and documentation, and their structural interrelations agree with those specified in the TEI DTDs.

28.2 Modifications to TEI SGML Declaration

The effective SGML declaration cannot be changed when using XML. When using SGML, the SGML declaration for TEI interchange documents may differ from that provided in TEI documentation in these ways:

- the CHARSET clause must be used to define the transmission character set (possibly in connection with the SHUNCHAR specification in the SYNTAX clause);
- the CAPACITY clause may be used to raise (but not lower) capacities;
- the SYNTAX clause may be used to define the SGML syntax used in the document. Notably:
  - the SHUNCHAR specification within the SYNTAX clause may be used to restrict the transmission character set;
  - the BASESET and DESCSET specifications within the SYNTAX clause must be used to describe the transmission character set;
  - the DELIM and NAMES specifications may be used to modify the SGML syntax. In particular, for consistency between SGML and XML, the NAMECASE specification may (perhaps even should) be used to set GENERAL NO, making SGML names case-sensitive; and the DELIM specification may (perhaps even should) be used to set NET to "/>", making SGML and XML EMPTY elements compatible;
- in the FEATURES clause, CONCUR may be set to NO if concurrent markup is not used in the document.

The following portions of the SGML declaration may not be modified in TEI interchange documents:

- the CAPACITY and QUANTITY values may be increased but not decreased;
- the SCOPE clause may not be changed;
- no new FEATURES may be turned on.

The SGML declaration for TEI-local-format documents may be modified without restriction. Some recommendations for usage are made in document TEI P1, but these recommendations are not normative.

28.3 Modifications to TEI Document Type Declarations

A TEI-conformant document (whether for local processing or for interchange) may make any change to the TEI-supplied document type declarations which is allowed by SGML and the controlling SGML declaration. All such changes should be made (or at least it must be possible to make them) within the SGML DTD subset, by defining TEI DTD modifications files as described in chapter 29 *Modifying and Customizing the TEI DTD*, and embedding the modification files within the DTD subset of a document whose document type declaration refers to the unmodified TEI main DTD, as in the following fragment:

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE TEI.2 PUBLIC "-//TEI P4//DTD Main Document Type//EN"
                       "tei2.dtd" [
  <!ENTITY % TEI.extensions.ent
    PUBLIC "-//ProjectName//ENTITIES Local modifications to TEI main DTD//EN"
           "project.ent">
  <!ENTITY % TEI.extensions.dtd
    PUBLIC "-//ProjectName//DTD Local element types for TEI main DTD//EN"
           "project.dtd">
]>
```

For reasons of convenience, it may be desirable in practice to create a *derived DTD* in which the local modifications have been integrated with the TEI main DTD in a single file. If such a one-file DTD is desired, it should be derived automatically from the TEI DTD and the local modifications files using appropriate software, rather than derived by hand-editing the TEI DTD files, as hand-editing increases the chances of error and inconsistency between the DTD modifications files and the one-file DTD. Documents in the TEI interchange format must use the form shown above, with a reference to the unmodified main TEI DTD and declarations of the local modifications files.

The following must remain true of the DTD after modification:

- the overall document must contain a single `<teiHeader>` element and a single `<text>` element, in that order; in the case of a corpus or collection the overall collection may have a `<teiHeader>` followed by a series of `<tei>` documents;
- the `<teiHeader>` element must include elements for:

  title statement the title of the machine-readable work and the names of those responsible for it;

  publication statement place and date of publication or distribution of the machine-readable document;

  source description bibliographic description of the copy text or source of the electronic text, including at least title, author, and edition.

A TEI-conformant document may be said to require *DTD extensions* if it:

- defines new elements;
- modifies the content model, declared content, or omissibility of any element;
- adds or modifies any attribute definitions;
- renames any elements, attributes, or attribute values;
- defines any new document types;
- declares any non-SGML, non-XML notations.

Without requiring DTD extensions, therefore, any TEI document may:

- define entities and parameter entities;
- include processing instructions and comments in its DTD subset.

For local-processing purposes, a TEI document may also, without requiring DTD extension:

- include link type declarations etc. in its SGML prolog;
- define and use short reference mapping in its DTD.

Note that TEI interchange documents may *not* include link type or short reference declarations because the SGML declaration for interchange does not allow them (nor does XML).

It is expected that the notion of DTD extension will be particularly useful in describing the classes of documents accepted or validated by software.

28.4 TEI Processing Model

This section is included for illustrative purposes only; it does not restrict the processing of TEI or other documents. It simply distinguishes a number of typical ways in which a project may choose to apply the TEI Guidelines to different kinds of processing.

28.4.1 Document Capture and Reclamation

First, data might be captured by keyboarding into a locally defined data capture format, or by scanning into a locally defined scanner-file format. From these initial forms, transducers might convert the files into a standard local storage format.

28.4.2 Local Storage Format and Application Software

The local storage format might be the input format of some application program used frequently by the project. In this case, transducers might be necessary to prepare data for processing by other applications. Alternatively, the local storage format might be independent of the formats used by application programs; transducers would be needed to prepare data for any processing. Such an independent format is useful if the local storage format needs to contain more information than any single application can conveniently handle.

The local storage format might be SGML- or XML-conformant without being TEI-conformant, e.g. because it uses local DTDs instead of the standard TEI DTDs, or because it uses a TEI local processing format. Local software may be used to validate a TEI local-processing format, to transduce documents into the input formats needed by applications, and when appropriate to transform documents into the TEI interchange format for exchange with other sites.

Finally, the TEI interchange format may be used as a local storage format. It is not expected that this will be a very common practice, since it is expected that most sites interested in TEI conformance will eventually acquire markup-aware software which have advantages of compactness or processing. In the absence of such software, however, some projects may find the TEI interchange format (or perhaps a restrictive variant of it) useful, because such a format can be relatively easy to parse with ad hoc software.

Whether the local storage format is strictly TEI conformant or not, it may follow TEI-recommended practice in its selection of textual features to be marked up, in its tag names, in its documentation practices, etc.

28.4.3 Enrichment and Other Processing

Over the course of the project, analysis and processing may result in interim results which may be incorporated into the locally stored copy of the text so that the interim results can be used in later processing. This process of enrichment can be carried out either by manual editing of the documents using conventional text editors, or by application programs.

28.4.4 Data Export

When a document is to be exchanged with another site using the TEI interchange format, it must first be transduced from the local storage format to TEI interchange form. If local documents are already TEI-conformant, this requires either no processing at all, or a relatively simple normalization which can be handled readily by the normalization facilities of most SGML parsers. If the local storage form is non-SGML conformant (and not XML), some transducer must be used to transform it into the TEI interchange format.

The TEI-interchange-format document must then be packed for shipping into the TEI packed interchange format, using a packing program. This program will gather the constituent parts (files) of a document into a single file, and ensure that the file contains no characters whose safe passage to the recipient of the data is endangered by the transmission path. If the ultimate recipient of the document is unknown, the set of safe characters is very small. The specific *transmission character set* however is independent of TEI conformance: any convenient set may be used where both parties agree. The packer will ensure that the transmission character set is properly identified.

28.4.5 Data Import

When a document is received from another site using the TEI packed interchange format, it must first be unpacked into a TEI interchange-format document in the local character set. It may then be necessary to *naturalize* it by translating it into the local storage format; if the local format is TEI- or SGML-conformant, no processing is needed (although some SGML processors may offer a facility for suppressing omissible markup).

28.4.6 TEI Conformance in the Processing Model

The notions of TEI interchange format and TEI packed interchange format are central to the exchange of documents using the TEI guidelines, whether the local storage format is TEI-conformant or not. The TEI interchange format and the TEI local-processing format may each be used as a local storage format, though the local storage format might well differ from either of these without materially affecting the use of TEI formats for interchange. The TEI interchange format being less flexible than the local-processing format, it is expected that sites using SGML-conformant software may use the latter, while sites without such software may prefer the former.

The notion of TEI recommended practice, it is hoped, will be relevant to decisions about what textual features should be recorded during data capture and will thus affect data-capture formats and the transducers which render captured files into the local storage format.

The TEI abstract structure may be useful in developing local non-SGML markup schemes for data capture or for processing with ad hoc application programs. It is strongly recommended that the TEI recommendations, as well as the TEI abstract structure, be used for such development as well.

28.5 Aspects of Conformance and Document Description

28.5.1 Character Sets

Neither the character sets used for local processing nor those used for transmission of interchange documents are restricted by the definition of TEI conformance. For local processing, users will typically use the system character set of their local system or some modification thereof. For exchange with known partners, users should choose any convenient character set; typically the most convenient is the set of all characters which:

- are transmitted successfully over the existing transmission link;
- occur in both sender's and receiver's local coded character sets.

For blind exchange with unknown partners a conservative choice of transmission set is needed to ensure that characters arrive correctly. How conservative the choice need be depends on the medium of transmission. The *ISO 646 subset* defined in section 4.1.3 *Characters and their encoding* remains the only guaranteed safe set of characters for the regional and international networks most widely used, although larger character sets are increasingly coming into use. This is largely because silent and not always reversible translation between character sets remains a feature of transmissions across current disparate networks. At the present time (1993) therefore, only the ISO 646 subset is recommended for fully blind interchange, although the full complement of ISO character sets may be used successfully in some subdomains.

In transmission by disk or tape, however, no silent translation is likely to occur, and so larger sets may be successfully used in blind interchange. The primary danger is a failure of software in the receiving machine to process the characters correctly; at this time (1991), ASCII or 94-character U.S. EBCDIC appear to represent the largest safe choices; other national character sets may of course be used if good internal documentation is also provided.

Note that the transmission character set does not associate specific binary encodings with the characters in the set. In the technical senses, it is a character set, not a *coded* character set. This means that a document may undergo various automatic translations from one coded character set to another (notably, in the case of transmission over international networks, from ASCII to EBCDIC or vice versa) without leaving the *transmission character set*.

For further discussion of the topics addressed in this section, reference should be made to chapter 30 *Rules for Interchange*.

28.5.2 SGML Declaration

The utility of various SGML constructs is discussed in section 2.2 of document TEI P1 version 1. The restrictions on SGML declarations and SGML usage in TEI interchange documents discussed above under 28.2 *Modifications to TEI SGML Declaration* are derived from that discussion. In the case of XML, no SGML declaration changes can be made.

No restrictions are made on SGML usage in the local processing format because such usage is best determined locally and has no impact on interchange.

28.5.3 Document Type Declaration

The document type declaration provided by the TEI, whether in its SGML or its XML form, is intended to cover as wide a variety of document types and processing needs as proved feasible. It is impossible, however, for any finite list of text elements to cover every need of textual research and processing. As a result, extension of the TEI DTD has no effect on strict TEI conformance, as long as certain restrictions are observed; these have the effect of ensuring that later users of a file can easily see what changes have been made to the DTDs and what the new tags are intended to mean.

The requirement that all new or modified tags be documented, however, is formally verifiable only to a limited extent. It is possible for a program to verify that for every tag introduced in a DTD modification, a corresponding record exists in a Tag Set Declaration. It is impossible, however, to verify using formal means that the entry in the tag set declaration makes sense. Purely formal conformance measures, therefore, must be supplemented with human inspection of the documentation.

The concept of DTD extension is introduced to allow the concise description of software which is designed to handle documents encoded using the published DTDs but which is not prepared to deal with tags not included there.[172]

All sections of the TEI DTD are subject to modification by the user, except that a documentary header must be provided and distinguished from the text itself, and that documentary header must include tagged elements identifying the document encoded and those responsible for the encoding. This ensures that all TEI-conformant documents will have at least this bare minimum of accompanying documentation.

28.5.4 Tag Usage and Feature Marking

The basic design principles of the TEI require the notion of *TEI conformance* to be applicable to existing electronic documents if they are translated into a proper format, without requiring the insertion of information not captured in the initial preparation of the text.[173]

At the same time, the TEI is charged with formulating advice to those engaged in the creation of new electronic texts and is required to distinguish what is actively recommended for general use from what is merely optional, provided for use by those engaged in a particular sort of work.

The notion of *TEI recommended practice* is introduced to allow the concise description of documents in which not only the requirements, but also the recommendations of the Guidelines are followed. It is hoped that while projects to convert existing electronic data may content themselves with achieving TEI conformance, projects to produce new electronic texts will produce documents following TEI recommended practice. To distinguish those projects which follow the TEI's recommendation to use SGML or XML markup from those which capture the same underlying textual features but do so using other markup, the notion of the *TEI abstract model* is introduced; it is this which another encoding can have in common with the TEI.

28.5.5 Non-SGML, non-XML Markup

In exchanging texts for use by others, the goal of an interchange format is to ensure that the information encoded in an electronic version of a text can be correctly understood and processed by the recipient as well as by the originator of the text. To assure the achievement of this goal, the definition offered here of TEI conformance restricts markup in TEI conformant documents to SGML or XML markup and to other properly declared notations. The latter are explicitly recommended for the encoding of tables, figures, etc. and so cannot reasonably be excluded. Since they do place a burden on the recipient for proper processing, the use of any such notation is defined to fall within the class of DTD extensions.

Because of the escape clause for graphics, etc., it is in principle possible to create a TEI conformant document by embedding a document using any arbitrary markup into a driver file containing a TEI header and a declaration for the appropriate markup as notation. Though it falls within the letter, such a practice falls outside the spirit of TEI-conformant document interchange.

[172] Some will regard such simplifications as useful ways of making it easier to develop software which accepts TEI-conformant documents; others will deplore the failure of such software to accept all TEI-conformant documents including those which extend the TEI DTD. In providing the notion of *DTD extension* for describing what documents are and are not accepted by such software, the TEI acts in the belief that such software will in fact be developed; it neither endorses nor deplores its construction or use.

[173] See document TEI PC P1 "The Preparation of Text Encoding Guidelines."