The Text Encoding Initiative: Flexible and Extensible Document Encoding

David T. Barnard and Nancy M. Ide

Technical Report 96-396 (ISSN 0836-0227-96-396)

Department of Computing and Information Science

Queen's University

Kingston, Ontario, Canada K7L 3N6

December 1995

Abstract

The Text Encoding Initiative is an international collaboration aimed at producing a common encoding scheme for complex texts. The diversity of the texts used by members of the communities served by the project led to a large specification, but the specification is structured to facilitate understanding and use. The requirement for generality is sometimes in tension with the requirement to handle specialized text types. The texts that are encoded often can be viewed or interpreted in several different ways. While many electronic documents can be encoded in very simple ways, some documents and some users will tax the limits of any fixed scheme, so a flexible extensible encoding is required to support research and to facilitate the reuse of texts.

1 Introduction

Computerized processing of structured documents is an important theme in information science [1, 13]. To take full advantage of progress in this field it is necessary to have convenient electronic representations of structured documents. The publication of *Guidelines for Electronic Text Encoding and Interchange* [14] by the Text Encoding Initiative (TEI) [16] was the result of several years of collaborative effort by members of the humanities and linguistics research and academic community in Europe, North America and Asia to derive a common encoding scheme for complex textual structures [11].

The TEI was conceived at a meeting held in 1987 at Vassar College in Poughkeepsie, New York. The Vassar meeting drew together researchers involved in the collection, organization, and analysis of electronic texts for linguistic and humanistic research. Some of those present were involved in research programs that included the production of large corpora of texts, while others had written innovative pieces of software to process electronic texts in various ways. There was a shared view that the enormous variety of mutually incomprehensible encoding schemes in use was a hindrance to research. To facilitate the interchange and reuse of texts, and to give guidance to those creating new electronic texts, some new scheme was required that could synthesize the best ideas and aspects of existing schemes.

The Association for Computers and the Humanities, the Association for Computational

Linguistics, and the Association for Literary and Linguistic Computing jointly sponsored a project to devise a new encoding scheme. Funding was obtained from the National Endowment for the Humanities (USA), Directorate XIII of the Commission of the European Communities, the Social Science and Humanities Research Council of Canada, the Andrew W. Mellon Foundation, and the various organizations—companies, government agencies, universities and institutes—where the participants worked.

The work of the project was carried out by committees dealing with specialized topics, and coordinated by a Steering Committee, a technical review committee, and two editors.

Information about the Initiative appears primarily in the Guidelines [14] and on the TEI's World Wide Web page [16]. The first three numbers of the 1995 volume of *Computers and the Humanities* deal with various aspects of the TEI; these have been reprinted in monograph form [12].

The new encoding scheme was required to:

- adequately represent all the textual features needed for research,
- be simple, clear and concrete,
- be easy for researchers to use without special-purpose software,
- allow the rigorous definition and efficient processing of texts,
- provide for user-defined extensions, and
- conform to existing and emerging standards.

The Standard Generalized Markup Language (SGML) [9, 10, 7] was a relatively new standard when the project began. Although SGML was thus not in wide use within the communities served by the project, SGML was chosen as the basis for the new encoding scheme. The TEI encoding scheme is a complex application of SGML. To promote acceptability of the scheme, it was necessary to ensure that:

- a common core of textual features could be easily shared,
- additional specialist features could be easily added,
- multiple parallel encodings of the same feature were possible,
- users could decide how rich the markup in a document would be, with a very small minimal requirement, and
- there should be adequate documentation of the text and the markup of any electronic text.

Here is a simple document encoded using the TEI scheme.

```
<!DOCTYPE tei.2 system ’tei2.dtd’ [
  <!ENTITY % TEI.prose ’INCLUDE’>
  <!ENTITY english.wsd system ’teien.wsd’ SUBDOC>
]>
<tei.2>
<teiHeader>
```

```
  <fileDesc>
    <titleStmt><title>Short document.</title>
    <publicationStmt><p>Unpublished.</p></publicationStmt>
    <sourceDesc><p>Electronic form is original.</p></sourceDesc>
  </fileDesc>
  <profileDesc>
    <langUsage><language id=eng wsd=english.wsd usage=’100’></langUsage>
  </profileDesc>
</teiHeader>
<text>
  <body>
    <p>A very short TEI document.</p>
  </body>
</text>
</tei.2>
```

The first line of this document identifies the document type by pointing to an external definition. The next two lines (contained within the square brackets on lines 1 and 4) contain definitions that override those in the external definition; their purpose here is to include the definitions of the tag set for prose documents, and to indicate that the document being encoded is written in English. From line 5 through to the end of the document, there are two parts. The *teiHeader* contains information describing the electronic document; this is akin to an extended bibliographic description. The *text* contains the encoded text itself;

in this case, there is a single line of text.

Realistic documents have these same constituent parts – a reference to the TEI definitions, some local selections of parts of those definitions, a header, and the text itself.

The remainder of this paper addresses three aspects of the TEI Guidelines. In the next section we describe the structure of the definitions–a flexible collection whose parts can be selected as required for a document. We then discuss the tension between generality and specificity in this definition that is intended to serve well for many document classes. The last aspect we address is the need to represent multiple views of a single electronic text.

2 The Structure of the TEI Specifications

From the beginning of the project it was clear that the specification produced by the TEI would be very large. It was also clear that most documents would only make use of some parts of the specification. It was thus necessary to structure the specifications in such a way that users could select those parts that apply to a specific document.

The mechanism for making this selection should be easy to understand and use, and it should not require any formalism or software beyond what is provided in SGML.

The final specification arranges features to be encoded into several categories. Each set of features is encoded using a set of SGML tags. The DTD is put together in such a way that sets of tags can be included in the DTD or excluded from it, and thus the tags are allowed in a document or prohibited, respectively [15].

These are *core* features that appear in all documents.

- provision for characters sets to be used in documents
- tags for specifying the TEI header

  (including the file description, encoding description)
- tags for specifying basic elements

  (including paragraphs, emphasis, quotations, list notes, simple pointers, references)
- simple text structure

  (including front and back matter, possibly nested divisions)

Two sets of tags cover the core features. The first contains the tags for specifying the TEI header. The second contains tags for basic elements like paragraphs. The defining aspect of the core is that it appears in all bases.

There is a *base tag set* for each major document category that can be encoded. These are:

- prose

  (contains only the core features)
- verse

  (structure within lines, rhyme, metrical structure)
- drama

  (cast lists, performances, stage directions)

- transcriptions of speech

  (utterances, pauses, temporal information)

- print dictionaries

  (structure of entries, grammatical and typographical information)

- terminological databases

  (terms and definitions)

Each document explicitly selects one of these base tag sets. Each of these bases includes the core features.

Some documents are more complex; they contain material that is from more than one of these bases. For example, a critical essay that included lengthy quotations from poems under discussion would require both prose and verse structures. There are mechanisms for combining base tag sets in a single document; the details can be found in the Guidelines.

Tag sets for *additional features* can be used with any of the base types. These are included as necessary, depending on the information to be represented int he encoded text. The additional tag sets include:

- linking, segmentation and alignment

  (extended pointers, segments, anchors)

- simple analytic mechanisms

  (linguistic segments, spans of text)

- feature structures

(designed for linguistic analysis but useful for recording any values of any named features detected in the text)

- certainty and responsibility

  (who tagged this and how certain is the encoding?)

- transcription of primary sources

  (deletions, illegibility)

- critical apparatus

  (sources for the text, attaching witnesses to spans of text)

- names and dates

- graphs, networks and trees

  (encoding data structures to represet information about text)

- tables, formulae and graphics

- language corpora

  (collections of text, sources)

For example, to encode a technical paper, one would choose the prose base. The base implicitly includes the core features. If the paper includes tables and names, the tag sets for these additional features can also be selected. Figure 1 shows these relationships.

Modular construction is a familiar concept. Procedural, functional and object oriented programming languages all have mechanisms for combining parts into a larger whole in a tb

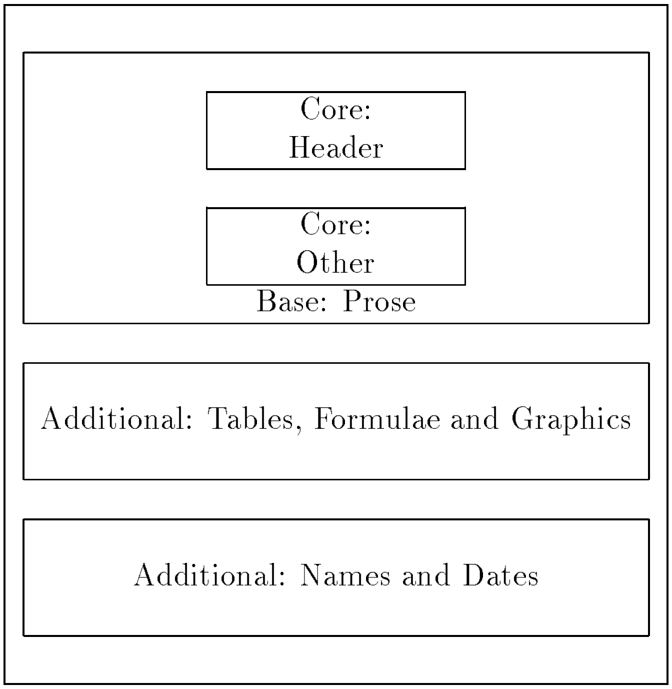
Core:
Header
Core:
Other
Base: Prose
Additional: Tables, Formulae and Graphics
Additional: Names and Dates

Figure 1: Tag Sets in the DTD

disciplined manner. In particular, object oriented notions of instantiation and inheritance support such constructions in a natural way. However, SGML was designed with very limited abstraction mechanisms; there is a SUBDOC construction, but it does not serve the purpose here.

To implement the modular structure that was required, the TEI DTD uses marked sections. There is a marked section for each of the tag sets. The section has a guard that is a parameter entity. A parameter entity is a string-valued variable that is defined in the DTD. When the value of the guarding entity is set to IGNORE, the marked section is not included as the DTD is parsed; when the value of the guarding entity is set to INCLUDE, the marked section is used. By default, all of the entities are set to IGNORE. The user explicitly resets the value of an entity to INCLUDE to select a tag set. Here is an example.

```
<!DOCTYPE TEI.2 system 'tei2.dtd' [
  <!--  base tag set                -->
  <!ENTITY % TEI.prose       'INCLUDE'>
  <!--  additional tag sets         -->
  <!ENTITY % TEI.names.dates 'INCLUDE'>
  <!ENTITY % TEI.figures     'INCLUDE'>
]>
```

This use of parameter entities, together with careful structuring of the DTD, does achieve a usable modularization. Bases and additional tag sets can be chosen as required.

However, SGML does not support modularization explicitly. Therefore, all of the tag sets must be named in the DTD whether they are used or not. Further, there is no capability for *scoping* as in block-structured programming languages. Thus all of the names in all tag sets must be distinct, and the use of tag names in the encoded document gives no indication of the set from which the name is drawn. This in turn has the result that the content of some tags is necessarily defined to allow far more variation than may be needed in a particular context.

3 Generality and Specificity

There is a tension between the need to encode general documents and the detailed requirements of encoding specific text types. Because it aims at maximal generality, the TEI necessarily takes its encoding solutions to the highest possible level of abstraction in order to accommodate a wide range of texts, disciplines, and applications, as well as different academic theories, languages, cultures, and historical timeframes. In addition, the TEI often provides multiple options for encoding the same phenomenon. As a result,

- TEI DTDs are over-generative–that is, they place very little restriction on where tags can appear relative to one another, in order to allow for even the most exotic of structures; and
- tags are often specified at a high level of abstraction, for example, the general <div> tag *versus* the more specific <chapter>.

The need to provide mechanisms which are maximally general and flexible is at times at odds with the provision of mechanisms which are most efficient or effective for a specific application or intended use. In particular, it interferes with the ability to *validate* the SGML document.

In SGML jargon, validation refers to the process by which software checks that the markup in a document conforms to the structural specifications given in the DTD–that is, that tags are properly nested, appear in the correct order, contain all required tags, etc.; that attributes appear when and only when they should, have valid values; etc. The ability to validate is important because it enables trapping errors during data capture. It also enables ensuring that the encoded text corresponds to the model given in the DTD, thus providing a possible means by which the adequacy of the model itself can be verified.

There is a tension between the generality of an encoding scheme and the ability to validate. Over-generative DTDs allow many tag sequences which, for any given text, are not valid. For example, to accommodate a wide range of dictionary structures, the TEI DTD for dictionaries allows most elements–orthographic form, pronunciation, grammatical information, etc.–to appear at any level inside an entry, and in any order. For any given dictionary, the structural rules are almost always much more constrained. For example, pronunciation may appear only at the highest level inside an entry and never within a given sense specification, and it may appear only following the orthographic form. However, with the TEI DTD it is not possible to validate that the tighter structural rules of this dictionary are followed.

The use of abstract, general tags also constrains the ability to validate. For example, the use of a general tag such as <div> to mark hierarchical divisions of a text (corresponding, for example, to book, chapter, section, etc.) disallows constraints on what can appear within a given text division. The <div> tag has to be defined to allow titles, paragraphs, etc. It is not possible to ensure that tighter structural constraints for a given book are observed, for example, that titles do not appear within chapters, or that a paragraph does not appear outside the chapter level, etc.

Because of the generality of the TEI scheme, it is likely that users will take extensive advantage of the TEI's mechanisms for modification and extension of the TEI DTDs to develop customized encoding formats in order to achieve better validation. The TEI is currently seeking collaboration with such users, in order to prevent the uncoordinated development of an increasing number of TEI "dialects" or sub-schemes.

4 Encoding Multiple Concurrent Views of a Text

A text encoded using SGML is most naturally to be thought of as a hierarchy: each component or element of the document contains others in a strictly non-overlapping, hierarchical manner. While this is a useful and powerful model, it does not represent all of the structures that documents of interest to users of the TEI Guidelines contain [3]. Some of the non-hierarchical structures of interest are these.

- One part of a document makes reference to another part of the document.
- One document makes reference to some part of another document.
- The target of a reference might be a point in the document, or a piece of text that is not necessarily co-extensive with some SGML element or elements.
- Parallel texts (such as a text and a translation of it into another language) with similar structures need to be aligned.
- Temporal correspondence of two texts spoken at the same time must be achieved.
- A conceptual part of the document might be formed from parts that are scattered throughout the physical document, such as speech interspersed with a running commentary.
- Correspondences between non-similar structures (such as two different parses of a natural language sentence) must be represented.
- An entire document might be represented with more than one structure.

Only the first category is directly representable in SGML, although some of the others can be represented in applications or extensions of SGML, such as HTML [6] or HyTime [8].

The last category is a generalization of the one that precedes it. Here we have more than one hierarchy, or parse tree, spanning a string. One example of this would be encoding verse drama so as to represent both the speakers and their speeches, and the verse lines and stanzas. Neither of these hierarchies contains the other.

SGML has a feature named CONCUR that is intended to deal with this case. However, there are several problems with it [5]. First, it is verbose: the amount of markup required to represent a set of features in a document is increased (by the addition of the specification of the DTD to which a tag pertains) if another set of features is represented in another view; it is, of course, to be expected that representing *additional* features requires more markup. Second, an entailment of the first drawback is that adding a new view to a document requires modification of the existing markup. Finally, CONCUR is not implemented by most SGML systems.

The TEI Guidelines include a number of techniques for solving linking and alignment problems like those described. Many of these techniques are based on the use of the ID and IDREF mechanism in SGML that allow a reference from one part of a document to another. In addition, attributes are placed on tags to specify semantics that indicate the nature of the links being built. The Guidelines also make use of *extended pointers*, that generalize the SGML mechanism by using string-valued attributes to encode additional information. Extended pointers can point to the same document in which they appear, or to another document (this is not possible in SGML). Extended pointers can be encoded in a simple query and pattern matching language, and thus be dynamically evaluated. Extended pointers can indicate a span of text starting at an offset from the beginning of a document component.

Some of the structures described above can be directly encoded with extended pointers, while some (such as aggregation of discontiguous parts) require structures that in turn can contain extended pointers.

Representing multiple parallel views can be done in several ways. One approach is to use milestones. These milestones are coordinate markers that can be placed in the text; separate structures containing pointers can refer to these milestones. Since there can be any number of pointers to a milestone, several parallel views of a text can be constructed int his way. For example, it is possible to represent parallel analyses of features of a text using several different collections of pointers to a set of coordinates in the text. These analytic structures do not need to be embedded in the running text, but can be placed outside of it.

For some frequently-occurring cases, there are more specific solutions provided by the Guidelines. The verse drama example given above is one such: there are specific tags for marking verse lines as incomplete, so that the information about lines can be fit into other hierarchies, such as that for speakers in a drama. Another example for which there are specific tags is the representation of pages and lines in specific editions of a text.

There are some weaknesses with the techniques used in the TEI Guidelines [4]. First, there are different ways of encoding the same notions in a TEI document; it would be unwise to strive for a minimal set of techniques (else we would program computers with the few instructions needed by Turing machines) but the diversity of the TEI scheme may make choosing the appropriate tags for an application more difficult than is necessary. Second, the TEI schemes rely on the interpretation of strings embedded in the attributes of SGML tags; application software that knows about these strings is not part of minimal SGML systems. Finally, some of the structures are complicated to represent; the alignment of a text with its translation is a simple concept, but requires many pointers to link the various components of the two texts (or parts of texts).

However, in spite of these weaknesses the approaches to encoding non-hierarchical structures, particularly multiple parallel views, used in the Guidelines are sufficient to represent a wide range of structures that are useful for a wide range of texts. While the structures used are more complex than what is available in SGML (because we have conventions about interpreting SGML attribute values), they are simpler than those represented in HyTime and more comprehensive in their semantics of relationships than those represented in HTML [2] or HyTime.

5 Implications and Lessons

The documents that are of interest to the community served by the TEI are complex documents, some of which contain many features that must be marked. Many of these documents contain features that are inherently non-hierarchical. While some of the documents of interest to researchers are particularly complex, the general observation that no simple specification will meet the needs of any large community of users remains true for other communities as well. The rapidly increasing complexity of HTML (HyperText Markup Language), the basis of the World Wide Web, is additional evidence that simple specifications are insufficient for large communities with interesting documents.

This means that a dynamic environment is needed for the specification of document encoding – an environment in which it is simple to encode simple structures, but in which more complicated structures can be encoded. In fact, it is important to have specifications that can be gracefully extended as new features of interest are identified.

The Text Encoding Initiative has produced a complex DTD that is modular and extensible. The DTD supports the encoding of a rich set of features in texts. The modularization mechanisms used in this DTD are effective, but somewhat clumsy. As SGML is used in other complex environments, other users will doubtless find a need for a modularization mechanism in the language itself, rather than being satisfied with adaptations of mechanisms designed for other purposes.

The TEI proposals are being used by many research projects, including these:

- The British National Corpus, a 100-million word balanced corpus of contemporary written and spoken British English, with word-by-word grammatical annotation.
- The Stockholm-Umea Corpus of modern Swedish, also with detailed linguistic annotation.
- The Map Task Corpus of ELSNET.
- The English-Norwegian Parallel Corpus Project, which is creating a grammatically

annotated parallel corpus.

- The European Corpus Initiative.
- The European Expert Advisory Groups on Language Engineering (EAGLES), who have in published papers endorsed the TEI as the basis for work on interchange formats for European research projects in the field of language engineering.
- The Multext and Multext-East Projects, which are using and extending the TEI Guidelines to encode linguistic corpora, including parallel corpora in thirteen European and Eastern European languages.
- the PAROLE project, involving a network of national corpus-building projects throughout Europe.
- Chadwyck-Healey, a European publisher of reference material and large collections of textual matter, which uses TEI encoding for their English Poetry Database, their Patrologia Latina database, and other full-text products now being planned.
- Cambridge University Press, who have announced a massive electronic edition of the Canterbury Tales, which will be encoded using TEI markup.
- the Lingua Parallel Concordancing Project, which aims at managing a multilingual corpus to ease students' and teachers' work in second language learning.
- The Memoria Project, a preparatory action funded under the MLAP programme involving the Bibliotheque Nationale de France, CAP, and two partners in the present

proposal (Oxford and Pisa), in the definition of the next generation of multimedia workstation.

- The Oxford Text Archive, which is participating in a UK Government funded Arts and Humanities Data Service to distribute TEI-conformant electronic texts over the Internet.
- Project Runeberg, a Swedish Internet-based project which is also disseminating electronic texts in TEI format.
- The Swedish National edition of August Strindberg at the Royal Institute of Technology in Stockholm.
- The NOLA (Networking of Literary Archives) project is making descriptions of the holdings of several archives in Europe available in a consistent way on the international network.
- The University of Michigan Humanities Text Initiative, along with the University of Michigan Press, is releasing a new textual resource, American Verse.
- Recently work has begun on developing an electronic version of the Thesaurus Linguae Latinae, or TLL. This work is being done under the auspices of the Consortium for Latin Lexicography (CLL).
- The Model Editions Partnership is a consortium of seven historical editions which has joined forces with leaders of the Text Encoding Initiative and the Center for

Electronic Text in the Humanities to develop editorial guidelines for publishing historical documents in electronic form and a series of demonstration models.

- The University of Pittsburgh Electronic Text Project is a research and development effort investigating the technology and policy issues involved in producing, collecting, and serving richly marked-up scholarly texts over the University and wide-area network.

- The Brown University Women Writers Project is creating and making accessible an electronic textbase of women's writing in English from 1330 to 1830.

With the successful production of the Guidelines, the TEI has entered a new phase. There remains technical work to be done in simplifying and rationalizing some of the existing features, and further tag set development is required in some areas. Customization of the Guidelines is required–as was anticipated from the beginning of the Initiative–for some specialized research projects. The Guidelines in their present form need to be validated.

All of these activities will occur as the Guidelines are used by projects such as those mentioned above. The TEI is actively developing its collaborative relationships with other projects to facilitate the validation and development of the Guidelines in a principled, consistent manner. Researchers interested in such collaborative efforts are encouraged to communicate with any member of the TEI Steering Committee [16].

Acknowledgement

The Text Encoding Initiative is the result of the collaborative work of many scholars and technical experts. The introductory material in the *Guidelines* names all of these persons.

References

[1] Jacques André, R. Furuta, and Vincent Quint. *Structured Documents.* The Cambridge Series on Electronic Publishing. Cambridge University Press, Cambridge, 1989.

[2] David T. Barnard, Lou Burnard, Steven DeRose, David Durand, and C.M. Sperberg-McQueen. Lessons for the World Wide Web from the Text Encoding Initiative. *World Wide Web Journal*, Issue One: Proceedings 4th International Conference on the World Wide Web, Boston:349–357, 1995.

[3] David T. Barnard, Lou Burnard, Jean-Pierre Gaspart, Lynne Price, C.M. Sperberg-McQueen, and Giovanni Battista Varile. Hierarchical Encoding of Text: Technical Problems and SGML Solutions. *Computers and the Humanities*, 29(3):211–231, 1995. Reprinted in [12].

[4] David T. Barnard, Lou Burnard, and C.M. Sperberg-McQueen. Lessons Learned From Using SGML in the Text Encoding Initiative. *Computer Standards and Interfaces*, accepted 1995.

[5] David T. Barnard, Ron Hayter, Maria Karababa, George Logan, and John McFadden. SGML-Based Markup for Literary Texts: Two Problems and Some Solutions. *Computers and the Humanities*, 22:265–276, 1988.

[6] Tim Berners-Lee and Dan Connolly. Hypertext Markup Language - 2.0, <*draft-ietf-html-spec-06.txt*>. Boston, HTML Working Group, 1995.

[7] Robin C. Cover. SGML Web Page. *http://www.sil.org/sgml/sgml..html*, 1994.

[8] Steven J. DeRose and David G. Durand. *Making Hypermedia Work: A User's Guide to HyTime.* Kluwer Academic Publishers, Boston/Dordrecht/London, 1994.

[9] ISO (International Organization for Standardization). ISO 8879-1986 (E) Information Processing–Text and Office Systems–Standard Generalized Markup Language (SGML). Geneva, International Organization for Standardization, 1986.

[10] Charles Goldfarb. *The SGML Handbook.* Oxford University Press, Oxford, 1990.

[11] Nancy M. Ide and C.M. Sperberg-McQueen. The Text Encoding Initiative: Its History, Goals, and Future Development. *Computers and the Humanities*, 29(1):5–15, 1995.

[12] Nancy M. Ide and Jean Veronis, editors. *The Text Encoding Initiative: Background and Context.* Kluwer Academic Publishers, Dordrecht, 1995.

[13] INRIA (Institute National de Recherche en Informatique et en Automatique). *Le traitement électronique du document.* ADBS Editions, Paris, 1995.

[14] C.M. Sperberg-McQueen and Lou Burnard. *Guidelines For Electronic Text Encoding and Interchange (TEI P3).* ACH-ACL-ALLC Text Encoding Initiative, Chicago and Oxford, 1994.

[15] C.M. Sperberg-McQueen and Lou Burnard. The Design of the TEI Encoding Scheme. *Computers and the Humanities*, 29(1):17–39, 1995.

[16] Text Encoding Initiative. Text Encoding Initiative Home Page. *http://www-tei.uic.edu/orgs/tei/*, 1995.