Core text identification for full-text databases

Lisa A Lehman,
Assistant Professor of Information Science,
Rasmuson Library,
University of Alaska Fairbanks

John A. Lehman,
Professor of Accounting and Information Systems,
Professor of International Business
School of Management,
University of Alaska Fairbanks

Abstract

This paper compares efforts in minimal and core level cataloging with those in full-text markup, and describes a set of standards for identifying documents in on-line text corpora.

Background

The experiences of researchers developing full-text databases over the past two or more decades has led first to a decision on the part of much of the academic community to standardize on SGML-based markup [Barnard, David, et al.], and second to the development of a common document type definition (DTD) for scholarly texts via the Text Encoding Initiative (TEI) project [Sperberg-McQueen and Burnard 1994]. The requirement that this common DTD meet the needs of all types of projects means that (at about 1,600 pages in the current edition) it is unwieldy for many uses. As a result, users have developed subsets, most notably bare bones TEI [Sperberg-McQueen, C. M 1995] and TEI-lite [Burnard and Sperberg-McQueen 1995].

Similarly in the library community, recent trends in declining support for academic and other public institutions coupled with significant increases in the volume of published materials have led to an increasing cataloging backlog and to the reexamination of rules for bibliographic cataloging, In addition to the full Anglo-American cataloging rules and the associated MARC standards, both a Minimal Level Record Standard [WLN 19951 and a proposed core standard [Cromwell 19941 have been developed.

This paper proposes that the Electronic Title Page entries of TEI-lite documents should be entered so as to be as compatible with the proposed core level cataloging standards as possible, but that the requirements of users of electronic databases will require additional information.

TEI lite

The full structure of the TEI-lite Electronic Title Page is illustrated in Appendix A; it has the following outline (from Burnard and Sperberg-McQueen 1995]):

20.1 The File Description
- 20.1.1 Title Statement
- 20.1.2 Edition Statement

20.1.3 Extent Statement
20.1.4 Publication Statement
20.1.5 Series and Notes Statements
20.1.6 Source Description

20.2 The Encoding Description

20.2.1 Project and Sampling Descriptions
20.2.2 Editorial Declarations
20.2.3 Tagging, Reference, and Classification Declarations

20.3 Profile Description

20.4 Revision Description

A minimal header has the following structure:

```
<teiHeader>
        <fileDesc>
                <titleStmt> ... </titleStmt>
                <publicationStmt> ... </publicationStmt>
                <sourceDesc> ... </sourceDesc>

      </fileDesc>
</telHeader>
```

The <fileDesc> section contains the bibliographic description of the electronic file. The <encodingDesc> section documents the relationship between an electronic text and the source(s) from which it was derived. The <profileDesc> section provides a detailed description of nonbibliographic aspects of a text, and the <revisionDesc> section describes the revision history for a file.

The File Description section which contains the bibliographic description of the from which it was derived. The <profileDesc> section provides a detailed description of nonbibliographic aspects of a text, and the <revisionDesc> section describes the revision history for a file.

The File Description section which contains the bibliographic description of the electronic file is made up of several parts.<tit]eStmt> provides the traditional user-oriented bibliographic data: title, author, editor, etc.<editionStmt> provides details on the edition of the original text which has been marked up. <extent> describes the size of the text. <pub I icati onStmt> provides information concerning the publication of the electronic text. <serie sS tmt> provides information concerning the series to which the electronic publication belongs (if any). <notesStmt> provides additional information about the text, and <sourceDesc> provides the bibliographic description of the copy text(s) from which the electronic text was made.

The<titieStmt>may contain <title>, <author>, <sponsor>, <funder>,<principal> (the principal researcher), and <respStmt>. <respStmt> contains a description of responsibilities and the name of the person attached to each, using the elements <resp> and <name>. <respStmt> and the associated elements <resp> and <name> are also used to provide details of the responsibility for the <edition> in the <editionStmt>

The <publicationStmt> may consist of a description of the publication circumstances, or may contain the elements <publisher> <distributor>, and .<authority>. All of these refer to the electronic edition, not the paper edition from which it may have been derived. They may use the following elements to describe the details of the electronic publisher: <pubPlace>, <address> <idno>, <availability>, and <date>

The <seriesStmt> may contain <title> or <respStmt> elements. The <notesStmt> contains one or more <note> elements.

The <sourceDesc> may be a general text description or may contain a formal bibliographic citation, using one of: <bibl> (for a loosely-structured bibliographic citation) or <biblFull> which contains all components of the TEI file description or <IistBibl> which is used for lists of bibliographic citations.

The <encodinaDesc> section documents the relationship between an electronic text and the source(s). It may be prose description or may be a structured description using the <projectDesc>, <samplingDecl>, <editorialDecl> (policies for dealing with topics such as correction, normalization, quotation, hyphenation, segmentation, and interpretation) < tagsDecl>, <refsDecl>, and <classDecl> elements, each of which includes a prose description of the purpose for which the file was prepared, the way texts were sampled to create the collection, details of editorial principles and practices applied during the encoding of a text, how the tags were applied, how references are constructed, and any classificatory codes used elsewhere in the text.

The <profileDesc> section provides a detailed description of non-bibliographic aspects of a text. It may contain <creation>, <IangUsage>, and <textClass> elements describing the creation of a text, sublanguages or dialects within the text, and terms of a standard classification scheme such as a thesaurus.

The <revisionDesc> section describes the revision history for a file. This is recorded as a sequence of <change> elements each of which contains <date>, <respStmt>, and <item> elements.

Minimal-level bibliographic records

While the fixed format of a MARC record is not directly compatible with a TEI Electronic Title Page, the content is compatible. Restricting ourselves to monographic records, a minimal level enhanced monographic record must contain the following items:

1xx Main entry
100 Personal Name

110 Corporate Name
120 Meeting Name
130 Uniform Title

245 Title Statement

250 Edition Statement

260 Publication. Distribution, etc.
Place of publication
Name of publisher
Date of publication

300 Physical Description
Extent

The minimum TEI Title page consistent with this format would thus be:

```
<teiHeader>
    <fileDesc>
        <titleStmt>
            <title>
            <author>
            <respStmt>
                <resp> compiled by </resp>
                <name> J
            </respStmt>
        </titleStmt>
        <extent>
        <publicationStmt>
                <publisher></publisher>
                <pubPl ace></pubPl ace>
                <date></date>
                <idno type=ISBN> </idno>
        </publicationStmt>
        <sourceDesc>
                <bibl>
                    <author>
                    <title></title>
                    <date></date>
                    <publisher></publisher>
                    <pubPlace></pubPlace>
                </bibl>
        </sourceDesc>
    </fileDesc>

<teiHeader>
```

Preliminary recommendations are that title pages for electronic documents should include at least the above superset of TEI-lite mandatory items, in addition to a <revisionDesc> section. As the proposed core standard for bibliographic citation becomes accepted in the international library community, the minimum TEI Title page should be expanded to meet the information requirements of this standard.

Bibliography

Barnard, David, et al., SGML-Based Markup for Literary Texts, Computers and the Humanities, 22 (1988): 265-76.

Burnard, Lou and Sperberg-McQueen, C. M., "TEI Lite: An Introduction to Text Encoding for Interchange" Document No: TEI U 5, June 1995, http://www.ulc.edu/orgs/tel/intros/telu5.tei.

Cromwell, Willy, "The Core Record: A New Bibliographic Standard," Library Resources and Technical Services 38(4) (1994) pp.415-424.

Goldfarb, Charles F.,The SGML Handbook. Oxford: Clarendon Press, 1990.

ISO (International Organization for Standardization). SO 8879-1986(E). Information processing --- Text and Office Systems --- Standard Generalized Markup Language (SGML). First edition -- 1986-10-15. [Geneva]: International Organization for Standardization, 1986.

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

ISO (International Organization for Standardization). ISO/TR 9573-1988(E). Information processing----SGML support facilities --- Techniques for using SGML. Final text of 1988-09-12.

Sperberg-Mc Queen, C. M., ans Burnard, Lou, TEI document P3, Guidelines for Electronic Text Encoding and Interchange, Chicago and Oxford in May 1994.

Sperberg-McQueen, C. M, Bare Bones TEI A Very Very Small Subset of the TEI Encoding Scheme, Document No. TEI U6 30 Aug 1994, rev. June 1995 http://www.ulc.edu/orgs/tei/intros/telu6.htmi

USMARC Format for Bibliographic Data. Prepared by Network Development and MARC Standards Office. Cataloging Distribution Service, Library of Congress, Washington, DC 1994. [Appendix A: National Level Record Requirements. Appendix B: Minimal LevelRecord Requirements]

Appendix A

TEI-lite title page sample

```
<!DOCTYPE tei.2 PUBLIC "-//TEI//DTD TEI Lite 1.0//EN"
[<!ENTITY amp
<TEI.2>
<TEIHEADER>
<FILEDESC>
<TITLESTMT>
  <TITLE>
title of text
</TITLE>
<author>
-Last name, first name, middle name, (birthdatedeathdate)
  </author>
<RESPSTMT>
<RESP>
  Describes responsibility for original data capture
  </RESP>
<NAME>-- whoever did the original data capture</NAME>
</RESPSTMT>
  <RESPSTMT>
<RESP>Converted to TEI form</RESP>
  <NAME>-- whoever converted it to TEI format </NAME>
</RESPSTMT>
</TITLESTMT>
<PUBLICATIONSTMT>
  <DISTRIBUTOR>the name of the project which is responsible for distribution
</DISTRIBUTOR>
  <IDNO type=???> 1 d-number</IDNO>
  <AVAILABILITY>restrictions on availability</AVAILABILITY>
  <DATE> I 997</DATE>
</PUBLICATIONSTMT>
<SOURCEDESC>
```