Three SGML metadata formats: TEI, EAD, and CIMI

A Study for BIBLINK Work Package 1.1

December 1996

Lou Burnard and Richard Light

Abstract

BIBLINK Work Package 1 (Study of Metadata) is intended to identify, describe, and compare current approaches to the encoding of metadata, with a view to making recommendations. Amongst the many different approaches currently in use, those based on the Standard Generalized Markup Lamguage (SGML: ISO 8879) appear to offer the widest range of features and the broadest potential for cross-sector applicability. This additional detailed study on three SGML-based formats was therefore commissioned to supplement the survey reported on in Work Package 1.

This report consists of a brief overview of each of the three schemes studied, some general discussion of the technical aspects of using SGML in a production environment, and a detailed feature by feature comparison of the three schemes with respect to a number of key attributes identified by the Biblink project. A bibliography with pointers to further reading on each of the three schemes is also provided.

Richard Light was responsible for the original draft of the report, and for the information on CIMI and EAD. Lou Burnard was responsible for final editing of the report and for the information on TEI and SGML. Thanks are due to John Perkins, Daniel Pitti, and Rachel Heery for helpful suggestions during the preparation of the report.

Contents

1 Introduction

Standard Generalized Markup Language (SGML) is a well-established international standard (ISO 8879:1986) for the description of mark-up languages. By *mark-up language* we mean here the formal system by which information or encoding is added to the electronic form of a document in order to represent its meaning, and hence to control its processing. Word-processors typically embed such information within a file using special proprietary control codes; database systems typically store such information externally in the form of a database schema. SGML allows for markup languages to be defined in a way which is independent of any particular device or application and thus allows for the interchange and long term conservation of richly structured electronic resources.

In this report, we consider one particular type of structured electronic document: the detailed bibliographic description, or *finding aid*. Such documents differ from most ordinary bibliographic records in their length and complexity, while at the same time differing from most ordinary textual documents in their highly structured nature. As such, they are particularly suited to an SGML-style encoding, since few other formats allow one to combine the rigour of a structured data record with the flexibility of a textual description.

We consider three specific sets of proposals for the use of SGML in the encoding of the kinds of metadata which are likely to be included in national bibliographies:

- TEI (Text Encoding Initiative) headers
- (EAD) Extended Archival Description
- CIMI (Consortium for the Interchange of Museum Information) records

It should be noted that SGML is increasingly widely deployed for this purpose. Amongst other SGML-based metadata proposals which we have not considered we may mention

- the ICPSR's SGML Codebook Initiative (to describe social science datasets)
- the FGDC (content standard for digital geospatial metadata)
- the CML (Chemical Markup Language) developed for description of chemical data sets.

These and other formats are described briefly in the Review of Metadata formats carried out as part of the DESIRE project (see Heery 1996).

2 Overview of the schemes studied

This section gives an introduction to the three SGML-based metadata formats described in this report. For each, we provide a brief history, a design overview, and an example of usage.

2.1 The Text Encoding Initiative

The Text Encoding Initiative (TEI) is an international research project, sponsored by three leading professional societies, with substantial international funding, from the Mellon foundation, the US National Endowment for the Humanities, the European Union Language Engineering Programme, and the Canadian Social Science and Humanities Research Council. Its primary goal was to define a set of recommendations for the encoding of literary and linguistic textual materials in electronic form, both in order to standardize existing work, and to facilitate the development of good practice in a rapidly developing field. The project began in the winter of 1987, and the most recent version of its chief deliverables, the two volumes of the *TEI Guidelines,* were published in May 1994 (Sperberg-McQueen 1994). Some indication of the wide range of work carried out within the TEI project is provided by the essays collected in Ide and Veronis 1996.; for a brief overview of the project's structure and organization, see Burnard 1994

The work of defining the TEI recommendations was carried out in a number of working groups and committee, with over a hundred volunteer contributors recruited from the international research community. Partly as a consequence of this large and varied user base, the *TEI Guidelines,* are extremely flexible: the end-result of the project was a modular, extensible, document type definition, combining a number of sets of element and attribute definitions, to be mixed and matched in a variety of ways according to the needs of particular communities. One of the most significant components of the TEI scheme is that defining a detailed bibliographic description known as the *TEI Header*.

2.1.1 The TEI Header

The TEI Header was defined in the first phase of the project, largely within the TEI working committee on Text Documentation, whose members included professional librarians and archivists as well as experts in markup. It was subsequently revised and expanded, with significant input from several TEI Working Groups, notably those concerned with the encoding of spoken language, and on the organization of language corpora. Its primary object is to address " the problems of describing an encoded work so that the text itself, its source, its encoding, and its revisions are all thoroughly documented. Such documentation is equally necessary for scholars using the texts, for software processing them, and for cataloguers in libraries and archives. Together these descriptions and declarations provide an electronic analogue to the title page attached to a printed work. They also constitute an equivalent for the content of the code books or introductory manuals customarily accompanying electronic data sets. " (TEI P3, p. 89). It is noteworthy that this 'electronic titlepage' is almost the only feature of the TEI encoding scheme which is mandatory.

It consists of the following four major sections:

- a file description, tagged <fileDesc>, containing a full bibliographical description of the computer file itself, from which a user of the text could derive a proper bibliographic citation, or which a librarian or archivist could use in creating a catalogue entry recording its presence within a library or archive. The file description also includes information about the source or sources from which the electronic text was derived.

- an encoding description, tagged <encodingDesc>, which describes the relationship between an electronic text and its source or sources. It allows for detailed description of whether (or how) the text was normalized during transcription, how the encoder resolved ambiguities in the source, what levels of encoding or analysis were applied, and similar matters.

- a text profile, tagged <profileDesc>, containing classificatory and contextual information about the text, such as its subject matter, the situation in which it was produced, the individuals described by or participating in producing it, and so forth.

- a revision history, tagged <revisionDesc>, which allows the encoder to provide a history of changes made during the development of the electronic text.

The structure of a TEI Header is fully detailed in the *Guidelines* and contains specific elements for a very wide range of elements (notably, almost all of those identified in the Survey of Libraries' Metadata Requirements reported in *Biblink study D1.1*, section 7). It should be stressed that this structure is architectural, rather than legislative: in other words, the TEI proposes a rich collection of metadata components, and a structure within which they can be expressed, and expanded. It provides little or no guidance as to the particular selection of such components which should be used by particular projects. Definition of such TEI Applications was consciously left to users of the scheme by its designers. Consequently, headers defined by different projects may vary widely. However, there are increasing signs of convergence amongst (for example) the practice of the growing number of electronic text centres and archives employing the TEI Header to document their holdings.

2.1.2 Sample TEI Headers

The following header is from the Victorian Women Writers Project at Indiana University:

```
<TEIHEADER><FILEDESC>
<TITLESTMT><TITLE>Liberty Lyrics (1895):
 a machine-readable transcription</TITLE>
<AUTHOR>Bevington, Louisa Sarah (Guggenberger) (1845-?)</AUTHOR>
<RESPSTMT><RESP>Transcribed and encoded by </RESP>
<NAME>Felix Jung</NAME></RESPSTMT>
<RESPSTMT><RESP>Edited by </RESP>
<NAME>Perry Willett</NAME></RESPSTMT></TITLESTMT>
<EXTENT>TEI formatted filesize uncompressed: 1426 bytes</EXTENT>
<PUBLICATIONSTMT>
<PUBLISHER>Library Electronic Text Resource Service (LETRS),
Indiana University</PUBLISHER>
<DATE>September 22, 1995</DATE>
<AVAILABILITY><P>&copy; 1995, The Trustees of Indiana University.
 Indiana University makes a claim of copyright only to original
contributions made by the Victorian Women Writers Project
participants and other members of the university community.
 Indiana University makes no claim of copyright to the original
text.  Permission is granted to download, transmit or otherwise
reproduce, distribute or display the contributions to this work
claimed by Indiana University for non-profit educational purposes, -

provided that this header is included in its entirety.
 For inquiries about commercial uses, please contact: <ADDRESS>
<ADDRLINE>Library Electronic Text Resource Service</ADDRLINE>
<ADDRLINE>Main Library</ADDRLINE>
<ADDRLINE>Indiana University</ADDRLINE>
<ADDRLINE>Bloomington, IN  47405</ADDRLINE>
<ADDRLINE>United States of America</ADDRLINE>
<ADDRLINE>Email: LETRS@indiana.edu</ADDRLINE></ADDRESS>
</P></AVAILABILITY>
</PUBLICATIONSTMT>
<SERIESSTMT>
<TITLE>Victorian Women Writers Project: an Electronic Collection</TITLE> -

<RESPSTMT><NAME>Perry Willett, </NAME>
<RESP>General Editor</RESP></RESPSTMT></SERIESSTMT>
<SOURCEDESC>
<BIBLFULL><TITLESTMT><TITLE>Liberty Lyrics </TITLE>
<RESPSTMT><RESP>by </RESP>
<NAME>L.S. Bevington</NAME></RESPSTMT></TITLESTMT>
<EXTENT>16 p.</EXTENT>
<PUBLICATIONSTMT>
```

```
<PUBLISHER>Printed and Published by James Tochatti,  </PUBLISHER>
<PUBLISHER>"Liberty" Press </PUBLISHER>
<PUBPLACE>London </PUBPLACE>
<DATE>1895</DATE>
</PUBLICATIONSTMT></BIBLFULL>
<P>The copy transcribed is from Michigan State University Libraries.</P> -

</SOURCEDESC>
</FILEDESC>
<ENCODINGDESC><EDITORIALDECL>
<P>All poems occur as DIV0.  Sonnets   are
attributed as "type=sonnets"; the rest are "type=poem".   All quotation
marks, hyphens,  dashes, apostrophes and colons have been transcribed
as entity references.  All < lg >  (line groups) are attributed as
cantos, stanzas, couplets, verse paragraphs, etc.  All poems  with
regularly indented lines use the attribute "rend" in  the < l > tag,
with the value   "indent1" for one tab stop, "indent2" for two tab
stops, etc.  All split lines are attributed as  "type=i" for the
initial portion, and "type=f" for the final portion.</P>
<P>All apostrophes and single right quotation marks are encoded as
&rsquo;.</P>
<P>Any hyphens occurring in line breaks have been
removed; all hyphens are encoded as  &hyphen; and em dashes as &mdash;.
 </P></EDITORIALDECL>
<TAGSDECL>
<TAGUSAGE GI="back" OCCURS="1"></TAGUSAGE>
<TAGUSAGE GI="body" OCCURS="1"></TAGUSAGE>
<TAGUSAGE GI="corr" OCCURS="4"></TAGUSAGE>
<TAGUSAGE GI="div" OCCURS="3"></TAGUSAGE>
<TAGUSAGE GI="div0" OCCURS="15"></TAGUSAGE>
<TAGUSAGE GI="div1" OCCURS="2"></TAGUSAGE>
<TAGUSAGE GI="docauthor" OCCURS="1"></TAGUSAGE>
<TAGUSAGE GI="docdate" OCCURS="1"></TAGUSAGE>
<TAGUSAGE GI="docimprint" OCCURS="1"></TAGUSAGE>
<TAGUSAGE GI="doctitle" OCCURS="1"></TAGUSAGE>
<TAGUSAGE GI="emph" OCCURS="15"></TAGUSAGE>
<TAGUSAGE GI="front" OCCURS="1"></TAGUSAGE>
<TAGUSAGE GI="head" OCCURS="18"></TAGUSAGE>
<TAGUSAGE GI="L" OCCURS="484"></TAGUSAGE>
<TAGUSAGE GI="lg" OCCURS="109"></TAGUSAGE>
<TAGUSAGE GI="p" OCCURS="7"></TAGUSAGE>
<TAGUSAGE GI="pb" OCCURS="14"></TAGUSAGE>
<TAGUSAGE GI="text" OCCURS="1"></TAGUSAGE>
<TAGUSAGE GI="titlepage" OCCURS="1"></TAGUSAGE>
```

```
<TAGUSAGE GI="titlepart" OCCURS="1"></TAGUSAGE>
<TAGUSAGE GI="titlestmt" OCCURS="2"></TAGUSAGE>
</TAGSDECL></ENCODINGDESC>
<REVISIONDESC>
<CHANGE><DATE>1995-06-30</DATE>
<RESPSTMT><NAME>Felix Jung, </NAME>
<RESP>editor.</RESP></RESPSTMT>
<ITEM>finished data entry, basic encoding and proofing</ITEM>
</CHANGE>
<CHANGE><DATE>1995-09-11</DATE>
<RESPSTMT><NAME>Perry Willett, </NAME>
<RESP>general editor.</RESP></RESPSTMT>
<ITEM>finished TEI-conformant encoding and final proofing</ITEM>
</CHANGE>
</REVISIONDESC>
</TEIHEADER>
```

This example demonstrates how traditional cataloguing (bibliographical) information, rights and permissions information, specific encoding details, and version information are readily combined in one descriptive framework. The same encoding framework also applies to the text itself, of course, since this is also encoded according to the TEI Guidelines. A computer application capable of handling such a resource is ipso facto capable of handling its associated metadata.

A second example TEI header is taken from one of the 4124 texts making up the British National Corpus (British National Corpus, 1994). In this project some of the tags proposed by the TEI have been renamed, and the flexibility of the scheme greatly curtailed. However, the basic structure remains the same.

```
<bncDoc id=BDHD0 n=ZIT04A>
<header type=text creator='dominic' status=new update=1994-04-19>
<fileDesc>
<titStmt>
<title>Minutes: Juniper Green Village Association — an electronic version
</title>
<respStmt><resp>Data capture</resp>
<name>W R Chambers</name></respStmt>
<respStmt><resp>Transcription</resp>
<name>Oxford University Press</name>
</respStmt>
<respStmt><resp>Encoding, storage and distribution</resp>
<name>Oxford University Computing Services</name>
</respStmt>
<respStmt><resp>Text enrichment</resp>
<name>Unit for Computer Research into the English Language,
University of Lancaster</name></respStmt>
</titStmt>
```

```
<ednStmt n=1>Automatically-generated header
</ednStmt>
<extent kb=188 words=12139></extent>
<pubStmt>
<respStmt><resp>Archive site</resp>
<name>Oxford University Computing Services</name>
</respStmt>
<address>
  13 Banbury Road, Oxford OX2 6NN U.K.
  Telephone:  +44 491 273280
  Facsimile:  +44 491 273275
  Internet mail: natcorp@ox.ac.uk
</address>
<idno type=bnc n=ZIT04A>
<avail region=world status=unknown>
<!- terms and conditions text summarized here ->
</avail>
</pubStmt>
<srcDesc><biblStr><monogr>
 <title>Minutes: Juniper Green Village Association</title>
</monogr></biblStr></srcDesc>
</fileDesc>
<encDesc>
<projDesc>
See project description in corpus header for
information about the British National Corpus
project.</projDesc>
<refsDecl>
Canonical references in the British National Corpus
are to text segment (&lt;s&gt;) elements, and
are constructed by taking the value of the n attribute
of the &lt;cdif&gt; element containing the target text,
and concatenating a dot separator, followed by the value
of the n attribute of the target &lt;s&gt element.
 </refsDecl></encDesc>
<profDesc><creation date='1990/1993'></creation>
<txtClass>
<catref target='wriAD920 wriASe4 wriATy3 wriAud3 wriDom4 wriLev1 wriMed4 -
wriPP920 wriSta1 wriTAS3 wriTim2'>
<keywords><term>minutes</term></keywords>
</txtClass></profDesc>
<revDesc><change n=1>
<date value=1993-12-22>1993-12-22</date>
<respStmt><resp>Unprocessed text received by OUCS</resp>
```

```
<name>fgk</name></respStmt>
</change>
<change n=2><date value=1994-02-07>1994-02-07</date>
<respStmt><resp>Processed text passed to UCREL</resp>
<name>gmb</name></respStmt>
</change>
<change n=3>
<date value=1994-03-25>1994-03-25</date>
<respStmt>
<resp>Segmented text received by OUCS</resp>
<name>bryant</name></respStmt>
</change>
<change n=4>
<date value=1994-04-19>1994-04-19</date>
<respStmt><resp>Initial accession to corpus</resp>
<name>dominic</name></respStmt>
</change></revDesc></header>
```

This example also demonstrates how the integration of the header and text within a single encoding framework can be beneficial. The <catRef> element in the header above specifies the descriptive (classificatory) categories applicable to the specific text to which it is attached, by reference only. A full definition for each category used in the corpus is supplied in an additional corpus header, which is prefixed to the whole corpus. Each individual text header references the parts of the corpus header which apply to it by means of TEI pointers, as in this case. This kind of linking mechanism is widely used within the TEI scheme, with obvious advantages of consistency and validation. As a further example, a <language> element can be given in the header to define each language used throughout a text. For a multilingual text, each portion in a given language will then reference the appropriate <language> element using its lang attribute. The lang attribute is applicable to *any* element in the TEI scheme, which makes it possible to indicate changes of language at any desired level of granularity, from sections or subsections down to individual words.

2.1.3 Other forms of metadata

The TEI scheme also proposes a number of mechanisms for the embedding of metadata within the body of a text (as distinct from in the header prefixed to one). These mechanisms vary widely in their technical sophistication and expressive power, since they are intended to cater for a wide range of analytic needs. At the simplest end of the scale, an <index> element is provided, which can be placed anywhere within a TEI text to generate an index-entry of some kind for this point in the text (this is functionally equivalent to the CIMI <topic> element discussed below); for more complex interpretative structures, the <interp> element may be used both to define an analysis, and to link it to a span of text; <interp> elements can also be grouped into hierarchically organized <interpGrp> elements.

The TEI also defines a specialized tag set for the encoding of analytic interpretations of any kind, based on the *feature structure* formalism. This powerful mechanism has great potential for the representation of formal systems of all kinds, but has not yet been widely implemented. (See further Langendoen and Simons 1995)

2.2 EAD

2.2.1 Background

The origins of the Encoded Archival Description (EAD) framework can be traced to a project initiated by the University of California, Berkeley, Library in 1993. The goal of the Berkeley project was to investigate the desirability and feasibility of developing a non-proprietary standard for machine-readable *finding aids*, that is, the inventories, registers, indexes, and other documents created by archives, libraries, museums, and manuscript repositories to support the use of their holdings. An additional motivation was the growing importance of networks as a means of gaining access to such information about holdings and the desire to extend the scope and richness of the information generally provided by traditional machine-readable cataloging (MARC) records.

The principles underlying the EAD are summarized as follows in an early definition of the project (Ann Arbor 1994)

- The information in a finding aid describes, controls, and provides access to other information, and thus is not an end in itself. Finding aids are not objects of study but rather tools leading to such objects.

- Although the encoding scheme does not define or prescribe intellectual content for finding aids, it does define content designation ... While there are certain elements that ought to appear in any finding aid, various intellectual and economic factors influence the depth and detail of analysis employed. Taking this into consideration, the encoding scheme is designed with a minimum of required elements, but allows for progressively more detailed and specific levels of description as desired.

- The standard preserves and enhances the current functionality of existing registers and inventories.

- The standard is intended to facilitate interchange and portability. It will increase the intelligibility of finding aids within and across institutions, permit the sharing of identical data in two or more finding aids, and assist in the creation of union databases. It will also ensure that machine-readable finding aids will endure changing hardware and software platforms because they will be based on a platform-independent standard.

- The needs of public users, curatorial and reference staff, and finding aid authors were given priority in the standard's design ... The designers sought to create a DTD that can be easily mastered and incorporated into routine finding aid production by staff possessing only a minimal knowledge of SGML.

These principles led to a design in which, at the most basic level, a finding aid document consists of two or three segments:

- a segment that provides information about the finding aid itself (its title, compiler, compilation date, etc.);
- an optional segment containing hand-generated 'front matter' (title page, acknowledgements or preface for the finding aid itself);
- a segment containing the actual finding aid, which provides information about a body of archival material (a collection, a record group, or a series).

Following the example of the Text Encoding Initiative (TEI), the group designated the segment about the finding aid itself the *header*, within which two types of information could be presented:

- hierarchically organized information which describes a unit of records or papers along with its component parts or divisions;
- adjunct information which may not directly describe records or papers but facilitates their use by researchers (e.g, a bibliography).

The hierarchy of descriptive information, reflecting archival principles of arrangement, generally begins with a summary of the whole and proceeds to delineation of the parts as a set of contextual views. Descriptions of the parts inherit information from descriptions of the whole.

2.2.2 EAD header

The mandatory EAD header is based on the *TEI header*. Its function is to provide a descriptive identification of the encoded archival description or finding aid. Its components are:

- ead id
- file description
- profile description
- revision description
- footer

This structure departs from the TEI Header in two minor respects:

The <eadid> is a formal, machine-processable name or address for a unique, authoritative <ead> instance. Its function as an internal reference identifier would normally be carried out by the global id attribute defined by the TEI. A more general <idno> element is also defined within the body of the TEI Header (see further below)

The <footer> is a note, disclaimer, warning, etc. that should be printed at the bottom of each page, displayed with each screen, etc. Again, there are several possible TEI equivalents, depending on the role or function of this note in a particular EAD application.

There are also, of course, minor differences of detail within the body of EAD header elements.

2.2.3 Sample EAD header

This is an example of an EAD header:

```
   <EADHEADER LANGENCODING="USMARC"
            FINDAIDSTATUS="EDITED-FULL-DRAFT">
 <EADID SYSTEMID="DLC" AUTHORITY="DLC"
        ENCODINGANALOG="856$f">jackson.sgm</EADID>
 <FILEDESC>
 <TITLESTMT>
 <TITLEPROPER>SHIRLEY JACKSON</TITLEPROPER>
 <SUBTITLE>A REGISTER OF HER PAPERS IN THE LIBRARY OF
 CONGRESS</SUBTITLE>
 <AUTHOR>
 <EXTPTR DISPLAYTYPE="PRESENT" ENTITYREF="lcseal">
 Prepared by Grover Batts
 <LB> Revised and expanded by Michael McElderry
 <LB>with the assistance of Scott McLemee
 </AUTHOR>
 </TITLESTMT>
 <PUBLICATIONSTMT>
 <DATE TYPE="finding aid created">1993</DATE>
 <PUBLISHER>Manuscript Division
 <LB> Library of Congress</PUBLISHER>
 <ADDRESS>
 <ADDRESSLINE>Washington, D.C. 20540-4860</ADDRESSLINE>
 </ADDRESS>
 </PUBLICATIONSTMT>
 <SERIESSTMT>
 <TITLEPROPER>Registers of Papers in the Manuscript Division of the -
Library of Congress</TITLEPROPER>
 </SERIESSTMT>
```

```
<NOTESTMT>
<NOTE><P>Edited Full Draft</P></NOTE>
</NOTESTMT>
</FILEDESC>

<PROFILEDESC>
<CREATION>Finding aid encoded by Mary Lacy, Manuscript
Division, Martha Anderson, National Digital Library, and
others, Library of Congress,
<DATE>1996</DATE>
</CREATION>
<LANGUSAGE>
<LANGUAGE>eng</LANGUAGE>
</LANGUSAGE>
</PROFILEDESC>

</EADHEADER>
```

2.2.4 EAD finding aid

The finding aid itself contains a mandatory <archdesc> (archival description) and an optional <add> (additional materials) element. The archival description contains a *descriptive identification*, containing key information such as creator, title and creation date, physical description (extent, object type, etc.), repository name and department, and notes. This descriptive identification may be followed by additional detailed information such as administrative information, biography or history of people or organizations involved, controlled access headings, or scope and content of the described material.

The archival description may also contain any number of *descriptions of subordinate components* (<dsc>s). These can be full descriptions, like the top-level description, or can take the form of lists or tabular displays of components.

2.3 CIMI records

The Consortium for the Computer Interchange of Museum Information (CIMI) works to promote the standards-based interchange of museum information. It is a membership organisation, supported by individual museums and museum organisations in North America and Europe. European membership includes the U.K. Museum Documentation Association, the Victoria and Albert Museum and the Aquarelle consortium.

CIMI adopted SGML as an interchange format in 1994, and has since used SGML in Project CHIO, an experimental distributed database of heterogeneous Folk Art resources (exhibition catalogues, object records, bibliographic references and authority files).

In the course of Project CHIO, CIMI developed an SGML application for textual museum information resources, which is applied to exhibition catalogues within CHIO. This application is based on the TEI and so shares its use of the TEI Header to describe the electronic text itself. Also, the encoding of the 'standard' features of the text (sections, headings, lists, bibliographic citations, etc.) follows normal TEI practice.

2.3.1 A TEI application

As noted above, the CIMI DTD was developed as a domain-specific application of the generic TEI framework. As such, it uses the standard features of the TEI Header to encode core metadata about each document. Particular emphasis is placed on bibliographic information and on access information and conditions (copyright statements, credit lines, etc.)

In addition to the standard TEI Header, the CIMI DTD introduces metadata concepts which apply within the document itself.

2.3.2 CIMI Access Points

A principal aim of Project CHIO is to provide online access to the relevant parts of documents in response to enquiries. These enquiries might come from the general museum-going public, or from museum professionals.

In order to do this, any aspects of relevance to potential queries need to be marked up. Also (less obviously) the *scope* of each search term needs to be made clear. If a section within a chapter describes a technique of interest (e.g. rug-hooking), then only that section should be returned to the searcher, not the whole chapter (and certainly not the whole book!).

A distinction was made between the main topic of discourse within a piece of prose (*'primary' access points*), and passing mentions of a topic (*'secondary' access points*). For example, a section of one book might be a biographical essay on Grandma Moses, whereas another book might have a passing mention of her name. Clearly, the biographical essay is likely to be much more valuable to a searcher, and so the entry 'person = *Grandma Moses*' would be encoded as a primary access point within that section.

The CIMI access points were developed through study of the questions asked of museums by researchers and the general public. The Categories for the Description of Works of ART (CDWA) were also used as background.

2.3.3 Topics

A particular requirement of the CIMI application was to associate these access points not only with discrete documents or document sections, but also with arbitrarily small *chunks* of text within the document, for example to answer questions of the type "Show me anything that talks about Grandma Moses" . This is achieved by the use of a special purpose <topic> element, whose attributes specify the CIMI access point concerned, and its particular value, and whose location (in SGML terms, its *parent*) specifies the document fragment concerned.

For example, both a paragraph and an entire article about Grandma Moses would include an element like the following:

```
<topic access-point="subject" value="Grandma Moses"></topic>
```

Thi is the method by which *primary access points* are encoded. The access-point attribute indicates which CHIO access point is involved. The value attribute contains the actual value of the topic. This is a completely general method, and can thus be used for a variety of designators. For example:

```
<topic access-point="identity-number" value="1969.11.1"></topic>
```

indicates a topic of "identity number = *1969.1.1*" .

2.3.4 Contexts

For general (public) access the *topic* mechanism is felt to be sufficient. However, CIMI's Project CHIO also aimed to support a more complex 'Museum Point of View', for which <topic>s alone were insufficient. Certain topic designators (date is an obvious example) have a meaning which is quite different in different contexts. There is also a frequent need to organize topics into a hierarchy. To provide this precision of retrieval, topics can be given a *context*. For example, this <context> element:

```
<context CHIO="creation"> ... </context>
```

applies the context of "creation" to anything inside it. So,

```
   <context CHIO="creation">
          <topic access-point="date" value="1860">
          </topic>
</context>
```

gives the date "1860" the context that it is a creation date, rather than (say) a date of birth or death.

context elements allow the primary access concepts to be qualified more exactly, and also allow them to be grouped together to form meta-records describing objects, people, places, events. etc. As well as <topic>s , a <context> elements can contain subordinate <context>s, thus permitting the definition of quite complex structures of metadata.

2.3.5 Sample CIMI meta-record

The CIMI access point mechanism is designed to be very flexible. By providing simple 'building blocks', the CIMI framework allows complex statements to be built up as required.

This meta-record describes an object entitled "Storm-tossed Frigate" , giving its artist, date of creation and current identity number:

```
  <topic access-point="object.work" value="Storm-tossed Frigate">
  <context CHIO="creation">
    <context CHIO="creator">
      <topic access-point="person" value="Chambers, Thomas" ROLE="artist">
    </context>
    <topic access-point="date-range" FROM="1825" TO="1874" EXACT="NONE">
    </topic>
  </context>
  <context CHIO="current-location">
    <topic access=point="identity-number" value="1969.11.1"></topic>
  </context>
</topic>
```

This set of data would typically be placed just inside a section of the text which describes that object, and so would associate the following index terms with that section:

access point	*value*	*context*
object/work	Storm-Tossed Frigate	
person	Chambers, Thomas	creation - creator (+ role = ‘artist’)
date range	1825 - 1874	creation [of object]
identity number	1969.11.1	current location [of object]

2.3.6 Inheritance

The CIMI approach depends on the concept of *inheritance* which is inherent to SGML. Each section within the document ‘inherits’ the topics assigned to the larger sections of which it forms a part. Thus if a whole book is ‘about’ Folk Art, then each chapter within it is also ‘about’ Folk Art. If one chapter within that book is ‘about’ weaving, then every section within that chapter is ‘about’ Folk Art *and* weaving, *and* any topics that are specific to that section:

```
<text><topic access-point="subject"
                     value="Folk Art"> [book-level index term]
 ...
  <div1><topic access-point="process.technique"
                     value="weaving"> [chapter-level]
   ...
    <div2><topic access-point="person"
                     value="Moses, Grandma"> [section-level]
     ...
```

In this case, the <div2> (section) is ‘about’ Folk Art *and* weaving *and* Grandma Moses.

EAD employs a similar convention in which each level of an archival description ‘inherits’ the information provided by its parent, higher-level, descriptions: “The <archDesc> element encompasses an unfolding hierarchy of descriptive information which, reflecting archival principles of arrangement, generally begins with a summary of the whole and proceeds to delineation of the parts. Descriptions of the parts inherit information from descriptions of the whole.” If all the text were stripped out of a CIMI-encoded document, the access point information that remained would have a similar structure to an EAD archival description.

2.3.7 Linking and naming

The CIMI DTD is meant to support a distributed resource, with contributing documents, object records and image files physically stored anywhere on the Internet. This led to linking conventions between e.g. documents and their associated images that were both robust and flexible. TEI *extended pointer* conventions are used to express complex links (e.g. to a specific passage in another document).

Long-lived links should not, as far as possible, be ‘hard-wired’ to a particular physical location. The location of resources is liable to change over time --- as has already happened within the lifetime of Project CHIO. Also CIMI wanted to facilitate the possibility of creating *mirror sites* with hyperlinks to a local copy of images etc.

In order to achieve a degree of insulation from the changes that occur in the siting and naming of Internet resources, CIMI recommends the use of *formal public identifiers* (FPIs) for external image files, documents, etc. A typical FPI has the form

```
-//XXX//YYY Name//ZZ
```

where ‘XXX’ identifies the naming authority, ‘YYY’ the kind of entity named (e.g. document type definition, entity set, document etc.), ‘Name’ is a human-readable long name for the entity, and ‘ZZ’ is the human language used for its definition. For example, the following FPI is defined for one of the TEI document type definitions:

```
-//TEI//DTD TEI Lite 1.0//EN
```

SGML’s use of entity references to identify system-specific references within a document means that only the entity definition needs to change when a different system identifier is needed. The use of FPIs within such entity definitions adds a further level of indirection, which can greatly increase portability and document independence. When FPIs are in use it is normal to define the mapping between an FPI and a real system identifier within a so-called *catalog* file (along with some other aspects of importance to an SGML application). The format for such catalogs is not defined by the SGML standard, but is currently in the process of definition by an influential group of SGML vendors and implementors called SGML Open.

Where it is not possible (or convenient) to create PUBLIC Identifiers for entity references, URLs are used as SYSTEM identifiers. This at least gives a form of reference which can be resolved directly by Web-aware software. Where appropriate (for example where referring to an image file), relative rather than absolute URLs are given, again with the intention of improving ‘portability’ of the reference.

CIMI is not alone in having to cope with the ‘link rot’ which is endemic to the current generation of distributed information systems. Its use of SGML will enable it to benefit from whatever solution is eventually found to this pervasive problem.

3 Technical context

This section discusses some technical aspects of SGML when used as a vehicle for metadata standards, specifically the role of a document type definition (DTD), the choice between descriptive and prescriptive styles,and the role played by SGML in quality control and resource discovery.

3.1 The role of a DTD

We noted above that SGML itself is not a convention for representing information, but a way of representing such conventions. SGML has little or nothing to say about how a document should be processed, what an application should do with it, or even what it means. It is not a protocol, in the sense that (say) Z39.50 is, nor is it a program. Different applications may use the information encoded in an SGML description in different ways, depending on their needs. A formatting application, for example, can choose to associate printing styles with particular elements, while a retrieval application can improve precision by searching only elements of a particular type, or within a particular context. What SGML offers is the way for such applications to interact with the same data in a mutually consistent and well-defined way.

The part of an SGML system which makes this inter-operability possible is the *Document Type Definition* or DTD. A DTD defines names, attributes, and co-occurrence restrictions for all the identifiable elements and entities used by a class of documents. It says nothing about their semantics: it is the role of supporting documentation or usage notes to do this.

In particular, a DTD says nothing about how a document should be rendered on paper or on a screen, any more than it does about which elements should be indexed for rapid retrieval. In order to render a document, therefore, an SGML application will need a specification additional to (or as a substitute for) the DTD: this is commonly known as a *stylesheet*. The recently-defined ISO Document Style Syntax and Specification Language (DSSSL, ISO/IEC 10179:1996) provides a standard for the definition of such specifications. Because this standard has only recently been adopted, most current SGML browsers and formatters tend to use their own stylesheet languages, but this is likely to change with the availability of more general purpose DSSSL-compliant formatters such as JADE.

3.2 Descriptive or prescriptive?

Some DTDs are purely 'descriptive': their goal is to specify all the elements that may appear in a large range of not particularly homogenous materials. Others are more 'prescriptive': their goal is to constrain as exactly as possible the contents of documents. Typically, a framework for encoding a range of existing material will aim to be descriptive, so that encoders can mark up "what is there" . On the other hand, DTDs designed to hold newly-created information can be as prescriptive as they like, since the information will be added along with the structure. A tightly-defined structure can actually be helpful, by reducing the number of choices that an encoder has to make. One advantage of the SGML approach (which supports unlimited repetition, recursion and field lengths) is that even a tightly-controlled structure can support large and complex documents.

Of the schemes studied here, the TEI is the most general (least prescriptive), in that it is intended to cater for the widest variety of documents. The EAD is more prescriptive, in that it is intended for use with a specific class of documents, all of the members of which share certain elements and are unlikely to include others. The CIMI metadata records are also more prescriptive, in that they represent a customization for a particular set of applications of the general framework defined by the TEI.

In any DTD there will frequently be a choice between an analysed set of elements and free text. For example, the TEI offers three levels of formality for recording bibliographic references, from free text with arbitrary subelements (<bibl>) through to a fully-structured reference (<biblFull>). Within the TEI Header, there is a choice between using analysed subelements and free text paragraphs in areas such as the Publication Statement. This flexibility means that one cannot be certain that (say) a Publisher Name will always be available in analysed form within the metadata since it is an optional sub-element.

Even if <publisher> were made a mandatory element (which could be done with a minor change to the DTD), the SGML standard provides no means of controlling its content. Any syntax conventions or vocabulary control must be supported by additional application-specific software.

3.3 Hyperlinks

One feature of SGML that is relevant to the study of metadata is its ability to represent hyperlinks, not just to another information resource but to a specific point within it. This linking can be used to point to non-SGML objects as well as to SGML-encoded documents. For example, an SGML hyperlink could point to an area within a graphic image, or to a range of frames on a video. Thus it is possible to set up SGML-encoded metadata that includes machine-processable links to single points and passages within a wide variety of resources.

Two hyperlinking schemes are deployed by the applications being studied. Both are system- and platform-independent. EAD uses HyTime, an International Standard (ISO 10744) application of SGML. TEI and CIMI use TEI *extended pointers*, a scheme provided as part of the TEI application. The scope of HyTime is larger, as is its complexity while the TEI scheme is both conceptually and computationally simpler. The designers of the two schemes have however gone to some lengths to maintain compatibility between them. Software support for both schemes is increasingly being provided within SGML-aware browsers.

3.4 SGML for quality control

What facilities exist for assuring the quality and consistency of SGML-encoded metadata? Conformance of documents to their DTD is checked by an *SGML parser*, a program which checks that documents match the tree structure defined by the DTD. It may also be configured to produce a normalized form of the document, in which the *element structure* is represented unambiguously. (An SGML document need not represent explicitly all of its structural markup: various types of *minimization*, such as the omission of contextually-determined tags, being permitted by the standard). The *Element Structure Information Set* (ESIS) output by such an SGML parser is an essential first step in the creation of an efficient general purpose SGML processing tool.

A parser checks only the syntactic validity of a document. As mentioned above, additional software is necessary to check the **semantic** correctness of the content, for example to check that only terms from a controlled vocabulary are employed. Such checking is inevitably application-specific, requiring the development of application-specific software, either from scratch or by customization of more generic systems. Provided that the DTD accurately reflects the structure of the metadata to be processed however, it will be possible to develop more powerful applications than could be developed in the absence of marked-up data. For application areas (such as museum and bibliographic data) where the information is inherently complex with many inter-relationships, the SGML approach is also likely to be simpler to implement than the use of relational databases.

A regularly updated (and expanding) list of SGML tools and products is maintained at http://www.falch.com/SGMLtools listing several hundred products, both commercial and public domain, categorized by function. Key functions include

- dtd maintenance and design
- document authoring
- document validation
- transformation
- formatting
- information retrieval
- document management

Semantic checking may be carried out during the process of document authoring, as a separate post-editing exercise, or both. SGML-aware application development systems such as sgmlc, Balise, or Omnimark can be used to construct modules for this purpose, either to run stand-alone, or integrated with authoring tools such as Author/Editor or WordPerfect.

Much, if not all, of the functionality of such integrated systems is also available in state of the art object-oriented document management systems such as Astoria, along with many other desirable features. However, object oriented technology has not yet reached the state of maturity where it can be considered a low-cost or wide-appeal solution.

For the immediate future it seems likely that hybrid document management systems will continue to dominate the market. The hybrid approach enables the system builder to use the known strengths of relational database technology (for example, with respect to document integrity, multiple access, etc.) in combination with the evident superiority of SGML as a document representation scheme, facilitating more sophisticated enquiry and retrieval facilities. This can be achieved, for example, by storing SGML objects as 'BLOB's within a relational database, or by modelling at least some part of the SGML conceptual schema directly in the relational database, with appropriate SGML documents being generated from the repository via SGML translation modules.

3.5 SGML in resource discovery

To use SGML metadata documents directly for resource discovery implies some sort of SGML-aware search engine, such as Open Text or BASIS. SGML-encoded metadata can always be converted to some other format, whether 'on the fly'or as a batch operation, but without an SGML-aware search engine it will be difficult to take full advantage of the rich structuring inherent in the SGML data.

As noted above, the absence of low-cost full- featured SGML-aware database or document management systems encourages a hybrid approach to document management. Management and control information is stored in a conventional relational DBMS, with all its advantages for integrity control and management, from which complex SGML structured documents are generated for loading into a static SGML-aware document retrieval system, with all its advantages for efficient searching in complex structures. Results obtained from the document retrieval system can then be easily down-translated into an interchange format conforming to some externally agreed protocol such as Z39.50, Dublin Core, or even MARC.

At the Oxford Text Archive, for example, all the information required for a TEI Header is stored locally in a conventional Microsoft Access database, from which TEI Headers are dynamically generated. It is planned to load the headers, along with the texts to which they refer, into a single federated document management system using Open Text software. This database will service all bibliographic enquiries about texts, as well as analyses of the texts themselves, via a single forms-based interface. This architecture will also permit dynamic extraction of metadata information in a variety of different formats for use by remote clients. The TEI Header is certainly rich enough to support clients requiring Dublin Core records (see further section 4.8.1 below); at the OTA, it is hoped to define the headers sufficiently accurately to permit also the automatic generation of basic level MARC catalogue records on demand.

Similarly, the Z39.50 protocol has successfully been used within Project CHIO to carry out searches on CIMI metadata encoded in SGML, although in this case the actual searching was carried out on a database derived from the SGML, not directly on the SGML itself. Also, the system used was unable to support the CIMI concepts of *context* or *inheritance*. The use of Z39.50 with SGML documents is still very much at an experimental stage.

4 Comparative analysis

In this section, we compare each of the three schemes under discussion with respect to the following criteria:

- user community
- control agency
- expression of metadata
- metadata concepts supported
- rules for formulation of content
- extensibility
- future development
- relationship to other schemes

4.1 User Community

In this section we briefly survey the current state of usage for each of the three formats under study, with a view to giving some indication of how widespread its deployment is at present.

4.1.1 TEI headers

As they are an integral part of the TEI scheme, TEI headers are routinely found in all TEI-encoded documents. TEI encoding is widely accepted in several parts of the research community, in particular amongst those engaged in the creation of electronic libraries and text centres, in electronic publishing (for example of scholarly critical editions), and in the creation of language corpora for use in Natural Language Processing. In the US, leading electronic library projects, such as those at the Universities of Virginia, Michigan, and Indiana all use TEI headers to document their holdings. Several major text creation projects (e.g. the Womens Writers Project at Brown University, the NEH-funded 'Model Editions Project' and many others are already committed to their use. It is hard to think of a major electronic text creation project in the academic context which would not at least start by first considering use of the TEI scheme.

In Europe, the TEI has been similarly successful, though the user profile has tended to be slightly different. For example, a number of highly visible commercial electronic publishing ventures (e.g. Chadwyck Healey's *English Poetry* and Cambridge University Press's *Chaucer's Wife of Bath's Tale*) have made use of it, and the TEI scheme has been mandated for use in corpus building and language engineering projects by a series of European expert groups.

Details of these and many other TEI applications are available from the TEI applications page, maintained by the project at http://tei-uic.edu/orgs/tei/apps/

For projects using the TEI Header, it is helpful to distinguish between its role as a quality control mechanism during resource creation and management on the one hand, and its role as a source of rich information for use in resource discovery on the other.

4.1.2 EAD

Although the EAD framework is still in beta-test form (with the first 'official' release scheduled for early 1997), it has already been widely adopted (at least in principle) within the US archives community. *"Within the first few months of alpha testing, scores of archives and libraries marked up selected finding aids." [see EAD background paper]*

The EAD format is likely also to be adopted as a standard by the archives community in the United Kingdom and may well emerge as an EU-wide standard. Repositories in the United Kingdom committed to EAD include Liverpool University, Glasgow University, and the University of Durham. The Public Record Office is currently conducting a pilot project with the aim of converting their listings to EAD, and there is growing interest from the British Library and NCA in developing EAD applications.

The Library of Congress has agreed to take on the task of maintaining the EAD. It is anticipated that the Society of American Archivists (SAA) will, at the appropriate time, organize an EAD advisory committee comprising representatives from the archival, library, and museum communities as well as acting as the maintenance agency for EAD.

4.1.3 CIMI records

The CIMI framework has been developed in the context of a recently-completed research project (Project CHIO), which explored the possibilities of using SGML and the Z39.50 search and retrieval protocol for museum information. So far, only CIMI members (mainly North American, but with some European representation) have actively used the framework, although the wider museum community is well aware of CIMI's work through a series of workshops and conference presentations.

Within Europe, the Aquarelle project has joined CIMI, and plans to develop the metadata aspect of its work.

It remains to be seen to what extent the CIMI framework will be adopted by the museum profession as a whole.

4.2 Control agency

In this section we briefly state the body or bodies responsible for the current and future states of the three formats under review.

4.2.1 TEI headers

Future development of the TEI is controlled by an Executive Committee. composed of representatives from the three sponsoring organizations and the two TEI editors. A larger Technical Review Committee was set up in 1996, which will take responsibility for the future development and maintenance of the TEI Guidelines. It is expected that a number of work groups will be set up to deal with specific development issues, the results of whose work will be ratified by the Technical Review Committee. This Committee will also take responsibility for the continued correction and maintenance of the Guidelines, in the light of experience gained during their use over the last couple of years. Membership and other administrative procedures of this Committee are similar to the ISO model, with particular domain-specific experts serving fixed renewable terms. (Further details are given in *Procedures for Maintenance and Extension of the TEI Guidelines* available from http://www-tei.uic.edu/orgs/tei/ed/edw48.tei)

The TEI has announced its intentions of setting up work groups to develop proposals on a number of specific topics during 1997. These include:

- extensions to the TEI Header e.g. for geospatial data, art historical information, manuscript description;
- further work on textual criticism, to include tags for analytic bibliography, codicology and physical description of primary sources;
- further work on encoding of historical dictionaries;
- further work on Writing System and character set problems.

4.2.2 EAD

The Library of Congress, Network Development/MARC Standards Office (ND/MSO) has formally agreed to serve as the maintenance agency for the EAD. As maintenance agency, LC will make the DTD and support documentation available and act as a clearinghouse for communications on the EAD, chiefly through the establishment of an electronic list and World Wide Web site.

The Society of American Archivists (SAA) will be responsible for ongoing supervision of the standard. It is anticipated that SAA will, at the appropriate time, organize an EAD advisory committee comprising representatives from the archival, library, and museum communities as well as the maintenance agency.

4.2.3 CIMI records

The CIMI Consortium itself acts as the control agency for the CIMI framework. If museums adopt the framework more widely, it is likely that one or more ‘official’ museum bodies such as the Museum Computer Network, the American Association of Museums or the U.K. Museum Documentation Association will become involved, to give the framework a more neutral support platform.

4.3 Expression of metadata

Both the TEI header and the CIMI access points are metadata whose primary purpose is to 'add value' to a specific SGML-encoded document. They might be termed *closely-coupled metadata*, in that the metadata forms part of the document itself. (The TEI header can be 'de-coupled' to form an *independent header*: the CIMI access points cannot.) However, both formats serve as useful examples of metadata techniques that can be applied within an SGML framework.

The EAD scheme is pure metadata. There is no presumption that the archive being described is in any particular format. An EAD description is an artefact that is new-minted with the specific purpose of acting as a finding aid.

4.3.1 TEI headers

The TEI design means that all the metadata is gathered up in the header, and is separate from the document. Such links as are present point *to* the header, either from the document (e.g. language) or from elsewhere in the header (e.g. classification system). This means that the document relies on the TEI header being present, but the header does not need the document in order to be meaningful.

The TEI scheme allows for classification of the content of a document to be accomplished at any degree of granularity, though it is easiest to do this at the text level using the <textClass> element within the Header's <profileDesc>. Finer-grained characterizations are however possible within the TEI scheme, using the decls attribute mechanism (which allows for any structural element to specify the particular set of declarations applicable to it, including its classification), or more generally by using the generic linking mechanisms. The CIMI scheme, as already noted, has similar flexibility, and was developed specifically to enable multiple levels of description.

4.3.2 EAD

All of the metadata describing an archival resource is stored in the <findaid> element. The EAD header (unlike the TEI Header) is *not* metadata, in that it describes the finding aid itself. It might be termed *meta-metadata*!

4.3.3 CIMI records

Metadata is stored partly as per TEI in the TEI header, and partly as <topic> and <context> elements nested just inside the element to which they apply. This second approach is unique to CIMI within the three schemes examined in this paper. It is an approach that might be applied elsewhere, if 'self-indexing documents' are required.

Embedding metadata within the body of a document has good and bad points. The positive aspects are:

- the metadata travels with the document, and is automatically available for setting up a searchable database.
- current SGML databases are more likely to be able to resolve queries based on this structure

The negative aspects are:

- this is not metadata in the usual sense
- the metadata cannot as readily be verified (for example against a controlled vocabulary)
- the metadata is not independent of the document it describes

It should however be noted that SGML-aware conversion software can easily extract this metadata and re-express it as an separate file containing *independent links (ilinks)* pointing to the correct place within the source document.

4.4 Metadata concepts supported

In this section we compare each of the three schemes under review in terms of the features each supports for the encoding of bibliographic description, access terms and conditions, and subject terms or classification.

4.4.1 TEI headers

Full, authoritative information on the TEI header is available in chapter 5 of the *TEI Guidelines* (Sperberg-McQueen 1994).

Bibliographic description: As previously noted, the <fileDesc> component of the TEI header is precisely designed to give full bibliographic information, and is "closely modelled on existing standards in library cataloguing" (op cit , p.93); "It is the intention of the developers… to ensure that the information required for a catalogue record be retrievable from the TEI header" (op cit p 137). Its component elements are taken more or less unchanged from analogous concepts in established bibliographic standards, chiefly the International Standard Book Description, and the Anglo American Cataloguing Rules.

Terms and conditions: The<availability> element "supplies information about the availability of a text, for example any restrictions on its use or distribution, its copyright status, etc." This element can take one of a small set of predefined values for a status attribute; it can also contain a complex set of rights and conditions presented as prose. Concepts such as price and licence information are not held in analysed form, but could be included as prose description or notes. Several elements relating to distribution (for example <availability> and <idno>) are represented within a repeatable <publicationStmt> element within the <fileDesc> element, and can thus be different for different publishers, distributors, etc. of a resource. The <publicationStmt> element also contains information such as the name and address of the distributor, publisher, or release authority, and any associated identifier such as an ISBN or URI.

Subject terms and classification: These are recorded within the <textClass> element, using one or more of three distinct methods. The <keyWords> element can be used to supply a list of descriptive keywords, either user-defined, or from a named authority such as Library of Congress Subject Headings. The <classCode> element can be used to specify a value from some pre-existing classification or taxonomy, such as UDC. The <catRef> element can be used to specify (by reference) a value from a classification or taxonomy supplied explicitly as a <classDecl> element elsewhere within the header.

Together, these three methods enable very detailed subject classification information to be added using a combination of currently well understood techniques. There is however, as usual, no recommendation as to how individual projects should choose amongst them.

4.4.2 EAD

Bibliographic information: This is not a prominent feature of EAD. The 'EAD header' provides a brief characterization of the finding aid itself, not the source archive. There is an <ADD> element, defined as follows: "adjunct to descriptive data. Optional. Provides for adjuncts to the descriptive information. This is information that will assist in the use of the archival material, but is not itself archival description." One of its components is an optional <bibliography> element, which in turn contains a <bibRef> for an actual citation. This element has a loose content model, analogous to the <bibl> element in TEI. It allows for a reasonable degree of bibliographic description, but does not attempt to enforce any particular level of description.

Terms and conditions: information is held within <accessRestrict> and <useRestrict> within the <adminInfo> element. <accessRestrict> contains information on gaining physical access to the material, while <useRestrict> describes what restrictions apply to the allowed use of the material once physical access has been gained. These are optional elements, that can appear within the description of a component of the archive at any level. No guidelines are offered on the preferred structure and content of these elements.

Subject terms and classification: these are comparatively detailed within the EAD scheme. A number of specific elements are defined, grouped within the <controlaccess> element, so named because it contains controlled access terms. Each of these has a specific tag such as:EAD

<corpName> An organization or group of people that is identified by a particular name and that acts, or may act, as an entity.

<geogName> A proper noun identifying the name of a geographical place, natural feature, or political jurisdiction.

<occupation> Occupations (including avocations) that are significantly reflected in the materials being described.

<subject> Specifies a subject term.

<genreForm> Types of material distinguished by intellectual content or physical characteristics.

These specific tags can be mixed in with a free text subject description, which rather complicates their usability for automatic topic extraction purposes.

4.4.3 CIMI records

Bibliographic information: CIMI records are conformant TEI documents, and can therefore use exactly the same components for detailed bibliobliogarphic description as discussed above for TEI.

Terms and conditions: Again, CIMI records can use the same components for this purpose as discussed above for TEI.

Subject terms and classification: Again, CIMI records could use the same components for this purpose as discussed above for TEI. However, the mechanisms available for linking text classification information to particular parts of a document were judged inadequate or too complex for use in CIMI. The CIMI application therefore extended the TEI scheme (using the TEI's built-in customization mechanism) to include two special-purpose metadata elements which can be anchored at any point within a document, with local scope. These elements (<topic> and <scope>) were discussed above, in section 2.3.

4.4.4 Summary of feature coverage

This table summarises how some broad metadata features are covered by the three schemes examined:

Attribute	TEI	EAD	CIMI
Bibliographic	<fileDesc>	<add> <bibliography> <bibref>	<fileDesc>
Terms and conditions	<publicationStmt> <availability>	<adminInfo> <accessRestrict> <useRestrict>	<publicationStmt> <availability>
Subject terms, classification	<profileDesc> <textClass>	<controlAccess>	<topic> optionally qualified by <context>

4.5 Rules for formulation of content

A consistent feature of the three SGML-based metadata schemes studied is that they are relatively relaxed about the way content is actually expressed. This can start at the structural level: a typical definition from the EAD Tag Library Description will be like the following, for <accessRestrict>: "... Contains: <head> (optional). followed by zero or as many as needed of the elements found in: Paragraph-level Elements."

In other words, access restrictions are to bedescribed in prose and no specific elements are provided to represent specific concepts relating to access. A similar situation is found in the other two schemes, although the TEI Header does provide more specific elements in some cases as an alternative to running prose.

Beyond this, the three schemes have the following to say about allowed content:

4.5.1 TEI headers

The guidelines for creation of *independent headers* within the TEI scheme give some indications of parts of the Header to which such constraints are likely to be of importance: "where there is a choice between a prose content model and one that contains a formal series of specialized elements, wherever possible and appropriate the specialized elements should be preferred to unstructured prose" Similarly, in the discussion of the <title>element: "The level attribute must be used to indicate whether this is the title of a book, journal, or series. It is highly recommended that the type attribute be used to distinguish the main title from subordinate, parallel, or other titles" However, even in the case of independent headers there is no indication that the syntax or vocabulary of entries should be constrained in any way.

4.5.2 EAD

EAD makes no recommendations for the actual syntax or vocabulary of any textual element. This is a typical instruction (for the <corpName> element): "This element contains text and may contain any elements found in the Linking and Formatting Elements, as many times as needed."

However, the instructions for entering the attributes which provide metadata about the corporation name are much more specific:EAD

role Used to specify the relation between the name and the item being described. The value supplied should be a word or phrase taken from the USMARC relator code list.

sources Used to indicate the source of the controlled vocabulary term contained in the element. Possible values are:

aat (Art and Architecture Thesaurus)

aacr2 (Anglo-American Cataloging rules, 2d ed., rev.)

dot (Dictionary of Occupational Titles)

4.5.3 CIMI records

CIMI shares the general TEI approach to content within the TEI Header. However, within its <topic> and <context> elements it provides more specific guidance, by proposing a set of specific values for the access-point attribute on the <topic> element, and the CHIO attribute on the <context> element. In both cases, the values are drawn from a closed list, itself compatible with the CIMI Z39.50 profile attribute set. Even here, however, there is an alternative mechanism for <context> which lets you declare other contexts as a value attribute, with a corresponding authority; this extensibility is further discussed in the next section.

The actual value of the <topic> is not constrained.

4.6 Extensibility

4.6.1 TEI headers

As noted above, the TEI scheme is designed to provide a framework which can be customized and extended to suit the user's exact requirements. Users can define their own custom tags, rename TEI elements to a form that is more acceptable within their community, define a new base structure for their information, undefine existing elements, modify content models etc. The published TEI Guidelines include examples of 'approved' extensions which were developed along with the Guidelines themselves.

Within this general framework, the TEI Header is rather less easily extended than other parts of the DTD. If additional concepts are required, the existing elements that are to contain them need to be 'undeclared', then re-declared with their amended content. This is rather less elegant than the standard methods for adding to a class of existing elements within the document itself, but functional.

In general, modifications to the header are associated with the selection of tagsets from the TEI scheme which imply that these predefined modifications will be needed. For example, one of the effects of selecting the TEI's predefined tag set for language corpora is to extend the TEI Header, by including tags for documenting demographic and other characteristics of the 'participants' in a written or spoken text. Another is to add a fourth way of classifying texts, in terms of their *situational parameters*.

Another example is provided by work currently in progress at the Bodleian library, where a rich set of descriptive tags for components of a traditional manuscript description has been defined, and grafted into the existing TEI Header structure, simply by redefining the <encodingDesc> element to include a new <mssDescription> element.

4.6.2 EAD

We cannot find any suggestion that the EAD has facilities that would allow users to extend the DTD.

4.6.3 CIMI records

Insofar as CIMI records use the TEI header, the above remarks apply. In the specific area of subject classification, both the <topic> and <context> elements have been designed to accommodate an open-ended set of descriptors. For example, <context> has a CHIO attribute which contains a fixed list of the CHIO 'context' access points, but it also contains a pair of attributes value and authority, which can be used together to provide a context taken from any authoritative source. This has already been used to encode museum concepts which do not happen to fall within the CHIO scheme:

```
<context value="measurements" authority="CDWA">
```

Thus two levels of extensibility are available. An open-ended range of classifications can be encoded using the existing framework. And it would always be possible for CIMI to extend its own fixed list of contexts and access points.

4.7 Future development path

4.7.1 TEI headers

As noted above, there have already been proposals for the extension of the TEI Header to handle the specific requirements of manuscript description: a recently-organized conference surveyed a range of activities in this area (see the Studley Manuscript Encoding Meeting). One of the workgroups to be set up by the newly chartered TEI Technical Review Committee will address this and related issues of extending the TEI Header in a controlled manner.

It also seems likely that the definition of a set of Guides to Good Practice in the application of the TEI Header to a range of materials, at least to the kinds of textual material held at electronic text centres, will consolidate existing and newly-emerging consensus on how best to make use of its flexibility.

4.7.2 EAD

The current version of EAD is undergoing beta-testing: presumably all development efforts will go into releasing the first 'official' version of EAD. It is probably too early to say what will happen subsequently, but existing use of the beta version (and commitments made to testers) already limit the ability to change the EAD scheme in a non-upwards-compatible manner.

4.7.3 CIMI records

The CIMI framework is less finalised than either TEI or the EAD scheme. CIMI will review the results of Project CHIO, and has an open mind as to how the format might develop. The lack of any significant deployment does give CIMI the flexibility to change its mind. It has a keen desire for interoperability, and plans to talk to both EAD and TEI about this.

4.8 Relationship to other metadata schemes

This section attempts to assess the overall position of these SGML-based metadata formats in the more general scheme of things. In particular. we examine the relationship between these formats and the Dublin Core, IAFA, and MARC.

4.8.1 Dublin Core

The Dublin Core is a currently much discussed set of metadata elements, which is increasingly regarded as providing a useful basis for general purpose resource discovery activities, particularly with networked resources. It has the merit of defining a small number of very generally applicable concepts, into which almost any more elaborated set of metadata concepts can readily be mapped. Examples of mappings between Dublin Core and EAD, and Dublin Core and GILS amongst others are available from Miller 1996; we list a similar 'cross-walk' for the schemes discussed here:

DC heading	TEI	EAD	CIMI
Subject	<textClass>	<controlAccess>	<topic> <context>
Title	<title>	<titleProper>	<title>
Author	<author>	<author>	<author>
Publisher	<publicationStmt>. <publisher>	<publisher>	<publicationStmt> <publisher>
OtherAgent	<sponsor> <funder> <principal> <respStmt> <resp> <editionStmt> <resp>		<sponsor> <funder> <principal>, <respStmt> <resp> <editionStmt> <resp>
Date	<publicationStmt> <date>		<publicationStmt> <date>
ObjectType	<textClass> <keywords>		<textClass> <keywords>
Form	SGML; implied	SGML	SGML
Identifier	<publicationStmt> <idno>		<publicationStmt> <idno>
Relation			
Source	<sourceDesc> <biblFull>		<sourceDesc> <biblFull>
Language	<langUsage> <language>		<langUsage> <language>
Coverage	<extent>		<extent>

4.8.2 MARC

Chapter 24 of the TEI *Guidelines* addresses specifically the question of mapping the components of the TEI <fileDesc> on to corresponding MARC fields. The mapping defined there implies that automatic conversion would be difficult, even though each data item would be in an appropriate MARC field or subfield. For example, there is no provision for the 'Main Entry' (or USMARC 1XX fields) in the TEI header. The main entry should be manually constructed by the cataloguer, using appropriate name authority control, and human intelligence to select from the information given in a TEI header the agency primarily responsible for the intellectual content of the work. There is an <author> tag, but the form of the name would have to be checked by a cataloguer before the main entry was constructed. Specific sets of values for the TEI defined attributes would need to be enforced before the TEI tags could reliably differentiate between name, conference, or title series; in their absence there is no simple mechanical method for determining which MARC tag (410, 411, etc.) should be used for series <title> and <idno>. Safe practice would be to load any series statements into 490 fields, and then to conduct authority work on those fields.

Since that date however, there has been considerable progress: for example, with the definition of the 720 generic author field, some of the above difficulties are removed. In a report commissioned by the Oxford Text Archive (Giordano 1996) a detailed mapping between the TEI Header and USMARC is proposed along with some more tightly specified cataloguing practices which together make feasible automatic loading of TEI Headers to USMARC records. The paper demonstrates that it is possible to create valid MARC records directly from SGML-encoded metadata, by defining a set of local practices and conventions in addition to the constraints enforced by the SGML document structure.

5 Conclusions

5.1 Using SGML to represent metadata

SGML has features which make it a very suitable format in which to hold metadata, which are intended to be long-lasting and system-independent. SGML is a well-established platform- and application- independent format, enforced and verifiable by an international standard, with an expanding user base, which is well respected and supported within the data processing industry. SGML-encoded metadata is likely to remain usable across different computing environments, without loss of information.

SGML is a powerful formalism, which can be used to model anything from very simple and constrained metadata (there is an SGML application for Dublin Core, for example) to rich and complex information structures, such as those made possible by all three of the schemes studied in this report. Metadata can be embedded in the document itself, as in integral TEI headers and CIMI data, or free-standing, as in independent TEI headers and EAD headers.

SGML can be used as an interchange format amongst non-SGML and SGML-aware software systems. In an extreme case, the mapping of `n` different formats each to and from SGML will be more cost-effective than the `n*n` mappings needed to support interoperability of `n` different formats. Even within a single institution, SGML can be adopted as a reference format, into and out of which system-specific representations of the metadata can be automatically translated, in situations where it is not convenient or cost-effective to use the SGML format directly.

This kind of hybrid approach is likely to become less attractive as the availability of low-cost SGML software tools increases. It should also be noted that the very richness and expressive power offered by SGML may pose problems in mapping into less sophisticated formats without information loss.

5.2 The three schemes studied

Each of the three schemes studied offers the possibility of an extremely rich set of metadata, way beyond the level, say, of Dublin Core. However, it is up to implementors to make effective use of these opportunities. There are very few mandatory elements in any of the schemes studied. Also, in the absence of syntax and vocabulary control, software cannot automatically extract or process useful metadata.

There is a major difference in the degree of generality between TEI and the other two schemes. As previously noted, the TEI Headers was originally designed to make feasible the recording of the information which a cataloguer would need to generate an ISBD-conformant catalogue record, but not necessarily without manual intervention and human intelligence. It was also designed to be extended for a wide range of less predictable applications, in fields where standardization is less well entrenched.

The CIMI and EAD schemes, designed for art historical and museum, and archival applications respectively, are more tightly customized to suit the needs of their respective communities. It is interesting to note how closely the basic structure and concepts of EAD overlap with those of the TEI, although the two were apparently developed independently. It is also noteworthy that the CIMI scheme was developed very specifically as an instance of the basic TEI architecture within an specific application field.

Even so, all three schemes remain very general. They provide the implementor with considerable flexibility — indeed, with quite enough rope to hang him or herself! Simpler, more constrained, solutions would not however provide anything like such a wide potential for expansion and customization to suit particular needs.

5.3 Use of schemes in combination

Another aspect of this flexibility worth comment is that the schemes need not be used in isolation of each other. For example, one might use the EAD scheme to describe individual archival holdings down to the item level and then use TEI headers to describe individual documents, where these were deemed of sufficient importance to warrant the effort. Equally, one could embed CIMI topic descriptors within an otherwise purely TEI conformant document.

The Bodleian Library at Oxford is currently experimenting with the first approach in its catalogue of Western manuscripts. The EAD is used to describe the collection itself in the same way as it has been used for a variety of other special collections. Access to the individual EAD records for resource discovery purposes is provided over the World Wide Web, using specially written software to translate between the HTML required for the Web and the more general SGML used by EAD. In addition, very detailed metadata about each manuscript is stored as a TEI header, using a set of Bodley-defined extensions to the standard header. Further details with examples are available at the web site http://www.bodley.ox.ac.uk/mss/.

Similarly, it is easy to imagine systems in which an SGML-encoded metadata scheme might effectively be used in conjunction with a non-SGML scheme.

6 References

[1] British National Corpus *The British National Corpus Home Page* available from http://info.ox.ac.uk/bnc/

[2] Burnard, Lou *The TEI: a brief overview* available from http://info.ox.ac.uk/ota/teij31

[3] Burnard, L. Miller, E, Quin, L. and Sperberg-McQueen, C.M. *A Syntax for Dublin Core Metadata*(1996) Unpublished paper available from http://users.ox.ac.uk/~lou/wip/metadata.syntax.html

[4] *Categories for Description of Works of Art* available from http://www.ahip.getty.edu/gii/cdwa/

[5] Ann Arbor 1994 *The Ann Arbor Accords: Principles and Criteria for an SGML Document Type Definition (DTD) for Finding Aids*

[6] Consortium for Interchange of Museum Information (CIMI) *The CIMI Home Page* available from http://www.cimi.org/.

[7] Consortium for Interchange of Museum Information (CIMI):*Project CHIO Home Page* available from http://www.cimi.org/CHIO.html

[8] Cover, Robin C. *The SGML web page* available from http://www.sil.org/sgml/sgml.html

[9] Giordano, Richard *Recommended Mappings OTA Header / USMARC / Dublin Core Elements.*Manchester: University of Manchester, Department of Computer Science, 1996Document ref: OTA.941025:01

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

[11] Heery, R. *Metadata: a survey of current resource description formats* available from http://www.ukoln.ac.uk/metadata/DESIRE/overview/

[12] Ide, N and Veronis, J. (eds) *Text Encoding Initiative: Background and Context* Kluwer, 1996

[13] International Organization for Standardization (ISO) *ISO/IEC 10744:1992 Hypermedia/Time-based structuring Language (HyTime)* Geneva: ISO/IEC, 1992.

[14] Langendoen, T.L. and Simons G. 'Rationale for the TEI Recommendations for Feature-structure Markup' (in Ide and Veronis 1996)

[15] Miller, Eric *Dublin Core Crosswalk* Available fromhttp://www.oclc.org:5046/~emiller/DC/crosswalk.html

[16] Miller, Paul: 'Metadata for the Masses'in *Ariadne*5 (1996). Available from http://ariadne.ukoln.ac.uk/ariadne/issue5/metadata-masses/

[17] Pitti, Daniel 'The Berkeley Finding Aid Project: Standards in Navigation'in *Filling the Pipeline and Paying the Piper* (Washington, D.C.: Association of Research Libraries, 1995), p. 161-166. See also http://lcweb.loc.gov/loc/standards/ead/eadback.html and http://sunsite.berkeley.edu/FindingAids/EAD/eadmodel.html

[18] Pitti, Daniel 'Encoding Standard for Electronic Aids: A Report by the Bentley for Encoded Archival Description Development' in *Archival Outlook* (Jan. 96, pp. 10-13.)

[19] Sperberg-McQueen, C.M. and Burnard, Lou *Guidelines for electronic text encoding and interchange (TEI P3)* Chicago and Oxford, ACH-ALLC-ACL Text Encoding Initiative, 1994. Also available from http://etext.virginia.edu/TEI/

[20] Text Encoding Initiative *The TEI Home Page* available from http://www-tei.uic.edu/orgs/tei/

[21] Text Encoding Initiative *The TEI Applications Page* available from http://www-tei.uic.edu/orgs/tei/apps/