A survey on metadata for describing and retrieving Internet resources

Ana Maria de Carvalho Moura[a], Maria Luiza Machado Campos[b] and Cassia Maria Barreto[a]

[a] *Instituto Militar de Engenharia – IME/RJ, Departamento de Engenharia de Sistemas, Praça General Tibúrcio, 80, Praia Vermelha, CEP 22290-270, Rio de Janeiro, RJ, Brazil*
E-mail: {anamoura,cassia}@ime.eb.br

[b] *Universidade Federal do Rio de Janeiro – UFRJ, Departamento de Ciência da Computação – IM/NCE, Cidade Universitária, Ilha do Governador, Caixa Postal 2324, CEP 20001-970, Rio de Janeiro, RJ, Brazil*
E-mail: mluiza@nce.ufrj.br

Metadata, or "information that makes data useful," have been considered by the database community basically as data in dictionaries used to control database management systems operations. More recently, metadata have been used to describe digital resources available across networks. This paper presents a survey of the state of the art concerning the use and importance of metadata, focusing on different standards and models found in the literature, describing how they serve as a basis for integrating heterogeneous resources on the Web and for developing more sophisticated search mechanisms.

1. Introduction

The World Wide Web makes up an enormous electronic information repository that is meant to be available to everyone over the many computer networks comprising the Internet. Yet, despite all technological advances which have made information retrieval faster and easier than ever before, this invaluable resource remains scattered, hard to locate, and difficult to integrate. Growth in size and heterogeneity represent challenges for designers of search systems. Finding and retrieving information on the Web is a process that relies upon indexing structures which span a vast number of resources in many different sites all over the world. However, the effectiveness of these indexes is highly dependent on the way these resources have been described by their providers.

In this context, the metadata concept plays an important role, not only because it allows a description of available resources, but also because it supports a variety of functions that users have come to expect from search mechanisms, such as: data localization, data assessment, data selection, and data retrieval. Besides, the use of metadata might also provide an important support to data administration tasks usually found in corporate environments. In that case, its main purpose is to describe the nature and possible interpretations concerning all data maintained by the enterprise, including those stored on corporate databases. In this way, besides promoting effective use of these resources it also supports control of data production and manipulation activities. Along such lines, a number of metadata standards have been proposed, some of them extending their reach to the Internet environment: Dublin Core, MARC (Machine Readable Catalogue), IAFA templates (Internet Anonymous Ftp Archive), TEI, and others.

This paper presents the state of the art of metadata in a context of distributed resources, giving a general overview of different formats or standards for describing distributed resources, the main mechanisms used to locate resources on the Web, and the most important architectural proposals to support metadata management for heterogeneous distributed information sources.

The rest of this paper is organized into seven major sections. Section 2 discusses the importance of metadata, showing how it is used to describe the information content in the context of different kinds of resources, from single files to CASE tools. Section 3 presents a brief overview of the most important metadata standards, including examples of their use. Section 4 presents a general classification for metadata components based on several other classification proposals. Section 5 analyzes search mechanisms on the Web and how they relate to metadata. Section 6 presents current research projects in the area, as well as some important initiatives of architecture design that accommodate different metadata standards. Finally, section 7 concludes with additional comments concerning the main issues discussed throughout this paper.

2. General use of metadata

Traditionally, metadata use has been associated with database management system catalogues or repository descriptors, providing information about both stored data and associated business processes. More recently, metadata have gained importance as an essential asset to support corporate data warehouse architectures, as well as to refer to any data used to aid the identification, description, and location of networked electronic resources.

In the following paragraphs we describe different metadata uses, according to the different information resources available.

2.1. Data storage and management systems

Only a small part of the entire amount of data effectively available in digital devices is stored in databases; most data still remain in single file systems. Metadata play an important role in providing for efficient control of these information resources. This, however, raises some important questions: which features are more appropriate to describe each specific resource? Which constitute the description's granularity level? How to store metadata and how to relate them to the data they refer to, taking into account format differences and device-storage incompatibilities?

Metadata have distinct uses, strictly related to the environment in which data is stored and organized. In the next paragraphs we analyze metadata use in different situations:

- **Files:** Only very few descriptions associated with conventional data files (type and size, for example) are generally available. The meaning and format of data are usually hidden on application programs that use these files. Usage of data stored in these files depends basically on the availability of their format description, which may be considered as a rudimentary example of metadata.
- **Databases and multimedia data:** In conventional database systems, metadata correspond to the information stored in the data dictionary or catalogue, including schema description and other control information. Object-Oriented Databases (OODB) include different abstractions to treat non-structured information such as sound, image, and video. In this context, the use of specific metadata appropriate for each media supports queries based on its contents such as the ViMod video model described in [Jain *et al.* 1994]. Metadata are also an important requirement to provide data integration across heterogeneous resources, such as federated databases [Seligman and Rosenthal 1996].
- **Data warehouses:** Data warehousing is the term given to the process of extracting data from different databases and applications of an organization and placing it into a single database for the purpose of decision support. Metadata are central to this environment to ascertain the origin of a particular data item, where it was moved to, and what changes happened to it. In addition, metadata can also be considered an important basis for interoperability, integrating schemas, views and supporting information exchange among homogeneous and/or heterogeneous environments and platforms.
- **Georeferential data:** Geographical Information Systems (GIS) provide the essential functionality to manage and control georeferential data, corresponding to a location on the earth's surface. Using metadata in this context becomes essential, because of the variety and complexity of data types involved as well as their diverse interrelationships. To emphasize the significance of that factor and the importance of using metadata in GIS, specific standard descriptions for georeferential data have been developed. FGDC (Content Standards for Digital Geospatial Metadata [FGDC]) and SAIF (Spatial Archive and Interchange Format) [SAIF] are standards which have been extensively used. The SEQUOIA 2000 [Anderson and Stonebraker 1994] project creates a schema description for digital satellite images based on the SAIF standard.
- **Heterogeneous and Interoperative Resources:** In new corporate environments, based on client/server computing with a diversity of distributed systems and applications, metadata emerge as a critical element to promote information exchange among tools from different vendors. MDIS (the Metadata Interchange Specification) [MDIS] represents an attempt to achieve a metadata interchange standard by defining an extensible mechanism to allow vendors to exchange common metadata as well as carry along "proprietary" metadata.
- **Hypermedia and WWW:** In this context, metadata are used to improve discovery and retrieval of hypermedia elements. This is the main focus of this paper, and this topic will be discussed in more detail in the following section.

2.2. Metadata on the Web

One of the biggest difficulties in developing search tools for the Internet comes from the vast heterogeneity of search spaces and protocols (such as HTTP, FTP, Gopher, WAIS), information resources, as well as access and indexing mechanisms.

Electronic documents in HTML format represent most of the objects accessed via WWW. Generally they do not have any extra information associated to improve their retrieval, depending exclusively on the keywords and full text indexing used by the available search tools. These tools have not addressed some problems such as: lack of precision in the result set; multimedia resources are not indexed; network overload due to the search strategies employed by these tools.

2.2.1. Requirements of metadata on the Web

There are three major aspects for the deployment of metadata: resource description, metadata production, and metadata use [Iannella and Waugh 1997].

The first aspect concerns the information that will be expressed by the metadata. The resource type and purpose of descriptors will determine this major question. The second aspect is production of metadata. Metadata represent a summary of data descriptions. When manually generated, it constitutes an expensive process. There is a general tendency to make this generation an automatic procedure, whenever possible. The third aspect concerns metadata use and access. It is especially important as a mechanism for resource location in distributed networks environments like the Internet. It contains embedded information like resource identification, subject and structure description, etc., all important for the resource access and availability.

Metadata provision poses some interesting challenges. The number of Internet resources is growing increasingly fast. These resources are dynamic by nature: new versions are introduced frequently and documents are often renamed or moved to other places. Very frequently, it is not obvious how to consider digital resources for indexing – for example, should a set of WWW pages relative to a research be described as a single unit or should each individual page be indexed separately?

Other fundamental issues related to metadata in this context are [Iannella and Waugh 1997]:

- Because so many different metadata standards exist, it is possible for a resource to be described by more than one set of metadata attributes. How should we deal with this situation when sets of different descriptors are involved?
- Extensions to existing metadata standards must be supported to accommodate local information and new types of resources.
- Internationalization of standards should be considered, as most of them are currently English-based.
- Metadata need to be closely associated to the resource they describe. Metadata consistency is essential for data use. Hence, metadata should be generated at the same time (or very soon after) the resource is created. They should be modified when the resource changes. However, some specific types of metadata, such as ratings for a movie or a critique of a document, may be generated separately.
- Metadata are data. Hence, they present storage and access problems, as well as similar difficulties for the correct interpretation of their content.

With the aim of getting better results when searching for information resources on the Web, the metadata community has put great effort in the last five years on discussing metadata standards, models, and protocols that could be integrated with the search mechanisms already in use. Some of them are discussed next.

3. Metadata standards and models

In this section we review some of the most important standards discussed in the literature. We have focused essentially on text-based ones, as an attempt to list and analyze a major part of the ongoing approaches. These text-based standards vary in complexity and on their primary goal: some have been largely employed for cataloguing bibliographical documents; others have been created for discovering resources on the Web, while others are geared to specific domains. The standards presented in this section are grouped according to their original goal.

3.1. Metadata standards for bibliographic cataloguing

The standards grouped under this category, namely the MARC model and its variants, have largely influenced other standards in different categories.

3.1.1. MARC

The MARC standard (Machine Readable Catalogue) [USMA 1996] was created in the late sixties in order to help classification services to enable an exchange of catalogue records among them. It has been used in library automation services, as the basis for manipulating library records for display and indexing. Variants of the MARC standard, such as UKMARC and USMARC, have emerged in response to existing format conflicts to attend to individual requirements of different libraries.

Although not originally conceived for describing network resources, the USMARC format became the basis of the OCLC Intercat Project [OCLC]. The project evaluated the feasibility of using the MARC standard associated with AACR2 (Anglo-American Cataloguing Rules) [AACR 1988] for describing and accessing Web resources of various types.

Example of formulation using OCLC MARC is given in figure 1.

It is difficult to modify this format, because it is a very structured standard. The creation of a new element, for example, requires an international consensus from the USMARC community. Bibliographic descriptions follow the rules set out in AACR2 and ISBD (International Standard for Bibliographic Description). As a matter of fact, the current USMARC model serves libraries and users poorly, most notably in cataloguing of microfilm reproductions of printed texts. This is due, in large part, to

```
084  813.412
100  Ana Maria Moura
245  00 Metadata Support to Internet Resources ...  $h [computer file] / $c Ana Maria Moura
256  Computer file
260  [s.1.] : $b Departamento de Engenharia de Sistemas, Instituto Militar de Engenharia, $c [n.d.]
520  An article on metadata standards and models ....
856  7 $2 http $u www.des.ime.eb.br   $d /~anamoura/publicacoes $f publicacoes.html $u
http:// www.des.ime.eb.br/~anamoura/publicacoes.html
```

Figure 1. An example of MARC encoding.

USMARC's intrinsically flat structure which requires fragmentation of hierarchically related components into separate discrete records. This record structure reflects the requirements of computers in the early 1970's when systems were highly attached to their data storage capacity.

Although MARC is clearly unsuitable to cope with the new operational requirements of emerging library systems, it will still remain in use for many years, because there are already billions of MARC records in online library systems. For this reason, mappings between MARC and other standards are a major trend nowadays [UKOL].

3.2. Metadata standards for text encoding and interchange

The standards presented in this category are strongly related to the use of SGML (Standard Generalized Marked Language), whose tags provide structure and access to bibliographic information for online systems.

3.2.1. Text Encoding Initiative (TEI) independent headers

TEI guidelines were published in [Burnard 1994]. Their main purpose was to define a set of generic rules for representing textual materials in electronic form, allowing resource interchange and reuse. The initial project aimed to develop guidelines to prepare and interchange electronic texts for scholarly research.

TEI guidelines define textual features in terms of SGML elements and attributes, grouped into tag sets. An element is a textual unit such as a paragraph. In the header, an element would be a unit such as title or author. An attribute gives information about a particular occurrence of an element and would be structured as an attribute-value pair.

A header describing the text should precede every TEI encoded text. The various elements in TEI are grouped into tag sets: core sets (elements required by all documents); base sets (element sets appropriate for a particular document class: verse, prose, drama, etc.); additional sets (elements for specialized treatment of text in different document classes); and auxiliary sets containing elements with specialized roles.

The TEI header is made up of:

- file description (the bibliographic characteristics of the document and its source),
- encoding description (editorial decisions concerning the treatment of the text and editorial process details),
- profile description (additional non-bibliographic information, such as: language, details of participants, subject classification, etc.),
- revision description (details of updates).

TEI guidelines are oriented to the description of objects and give no consideration for service descriptions. Independent headers can be manipulated, searched, and retrieved by any software supporting SGML. Due to the complexity of dealing with TEI code, a simplified version of TEI has been created – TEI lite, which the Oxford Text Archive has used for encoding texts [Heery 1996a]. The Electronic Text Center of the University of Virginia Library has also used TEI lite to produce HTML documents automatically converted from TEI encoding, providing the addition of TEI descriptive terms to the text of an image file [Seaman 1994].

```
<teiHeader>
    <fileDesc>
        <titleStmt>
            <title> Metadata Support to Internet Resources ... </title>
            <respStmt><resp>compiled by</resp>
                <name>Ana Maria  Moura</name>
            </respStmt>
        </titlestmt>
        <publicationStmt>
            <distributor>Instituto Militar de Engenharia</distributor>
            <sourceDesc>
                <bibl> Master Thesis Monography </bibl>
            </sourceDesc>
    </fileDesc>
<teiHeader>
```

Figure 2. An example of TEI encoding.

Example of formulation using TEI (supplying only the minimal level of encoding required) [Burnard 1994] is given in figure 2.

This standard has a flexible encoding: only <titleStmt>, <publicationStmt> and <sourceDesc> tags are mandatory. The reliability of TEI records for information retrieval may be variable, because of TEI's flexibility in encoding. This procedure requires less training than USMARC, but it does not provide electronic address information for network resources within the header, although the <noteStmt> tag may be used for this purpose.

TEI guidelines evidence a limitation of MARC's ability to structure non-bibliographic information (such as hierarchy and collections), which can be used to evaluate electronic documents. TEI headers provide full bibliographic information, like MARC, but they add non bibliographic documentary information supporting more detailed analysis of electronic text.

3.2.2. Encoding Archival Description (EAD)

The EAD project has been in development since 1993 at the University of California in Berkeley. Its main goal is to develop a non-proprietary standard for machine-readable finding aids [Swetland 1996]. Finding aid tools such as inventories, registers, and indexes are created by archives, libraries, and museums to enhance their holdings' usage, providing abilities to control, navigate and describe archival materials (collections, items, etc.). The EAD encoding scheme is based on SGML and has been developed to describe textual and electronic documents, visual resources, and sound recordings. Starting from predefined requirements and focusing on inventories and register abilities, the EAD DTD (Data Type Definition) provides a hierarchy of descriptive information: the archival material is primarily described as a whole in a summary. Its component-part descriptions inherit information from this description,

```
<ead>
 <eadheader>  (contains descriptive and declarative information about the finding aid itself)
   <eadid>98765432</eadid>
   <filedesc>
       <titlestmt>
         <titleproper> Metadata Support to Internet Resources Description and Retrieval</titleproper>
        </titlestmt>
    </filedesc>
</eadheader>
<frontmatter>   (contains introductory materials or any front matter necessary for the formal  publication of the
encoded finding aid)
   <titlepage>
       <titleproper> Metadata Support to Internet Resources Description and Retrieval</titleproper>
       <publisher> System Engineering Department </publisher>
    </titlepage>
</frontmatter>
<findaid>  (contains administrative and intellectual information about the material being  described
               by the encoded finding aid)
    <archdesc audience= "external"  level="collection">
     <did>
         <unittitle>Technical Reports on Computer Science
         <unitdate>1997</unitdate></unittitle>
        <physdesc><extent>3 linear feet.</extent></physdesc>
     </did>
     <admininfo>
          <acqinfo>
              <p> work produced from Master Thesis </p>
          </acqinfo>
<p>The copyright interests in the Moura, Ana Maria & Hess Lilia.....</p>
          </accessrestrict>
     </admininfo>
      ........
 </findaid>
   ............
</ead>
```

Figure 3. An example of EAD encoding.

preserving the hierarchical relationships that exist between levels of descriptions and reflecting archival principles of arrangement. Its hierarchical organization provides information about finding aids in the EAD header segment. Another segment contains information related to the body of archival material. Adjunct information (such as bibliographic references) may also be added and a title page provides information about the repository identification or the type of finding aids.

The prototype of EAD DTD is currently available at two sites: the University of California in Berkeley and at the Library of Congress. Due to space limitations, in figure 3 we present only a small part from an example using EAD, based on formulations that are referenced in [Swetland 1996].

Although TEI and EAD have different goals, EAD developers have started considering TEI guidelines. TEI has been designed specifically to encode literary texts while EAD focuses on types of descriptive metadata that archival finding aids represent. TEI header structure, element names and attributes are used wherever possible. EAD provides a flexible and detailed data structure for archival description similar to MARC. However, it does not require authoritative forms of content for any of its elements, a major drawback for information exchange and retrieval. While MARC records provide summary description and access, EAD is intended to provide detailed description and access to archives and manuscript library collections. It accommodates registers and inventories of any length, describing the full range of archival holdings in various media.

3.3. Metadata standards for discovering resources on the Web

In this category, metadata are used in the context of gatherer programs (Harvest, for example) to support site administrators in describing various resources stored in their servers. Indexing data gathered by automatic tools have no standard format and lack any kind of explicit semantics. An effective interchange of updated and useful information among indexers becomes very difficult. In the standards presented in this section metadata are used by the gatherer and distributed, utilizing indexing services in order to provide a better recall of results collected on the Web. These mechanisms extract metadata from META tags contained in the HTML documents.

Template-Type:	DOCUMENT
Template-Version:	1
Title:	Metadata Support to Internet Resources ...
Author-Name:	Ana Maria Moura
Author-Email:	anamoura@ime.eb.br
Author-Work-Phone:	55-21-295 3232
Last-Revision-Date:	15 March 1998
Description:	This paper describes the most important metadata standards and models.
Publisher-Organization Name:	Instituto Militar de Engenharia
Keywords:	metadata, resource description
Format-v0:	Application/zip
URI-v0:	http://www.des.ime.eb.br/~anamoura/meta97.zip
Size-v0:	4 pages
Language-v0:	Portuguese
Format-v1:	Application/postscript
URI-v1:	http://www.des.ime.eb.br/~anamoura/meta97.ps
Size-v1:	5 pages
Language-v1:	English

Figure 4. An example of IAFA templates.

3.3.1. IAFA/WHOIS++

IAFA (Internet Anonymous Ftp Archive) [Deutsch *et al.* 1995] templates were designed by the working group of the IETF (Internet Engineering Task Force) to facilitate effective access to ftp (file transfer protocol) archives by means of describing the contents and services available in the archive. The main goal was to construct a record format which could be used by ftp-archive administrators to describe the various resources available from their own archives, including: images, documents, sounds; services such as mailing-list archives, usenet archives, datasets, and software packages.

The original IAFA template format has been developed for use with the Whois++ protocol, a directory-service software, which allows search and retrieval of centrally-created databases, offering also the possibility of searching across multiple databases [Falstrom *et al.* 1997a]. The Whois++ service uses templates (based on IAFA models) to provide structured information within its databases, as collections of data elements that are simple atributes [Falstrom *et al.* 1997b].

ALIWEB [ALIWEB] was the first system to implement IAFA templates. Other implementations following this metadata standard can be mentioned: DIGGER [DIGGER]; projects SOSIG (Social Science Information Gateway) [SOSIG] and OMNI (Medical Information Gateway) [OMNI]; HENSA Unix Archive [Beckett 1995], which uses IAFA templates for a database containing information on parallel computing; and NetEc [NetEc] which provides a database of resources in economics.

An example of a formulation using IAFA templates is given in figure 4. It is important to notice that the same intellectual content can be associated with several formats such as PostScript, HTML, etc. The variant fields URI-vi, format-vi, size-vi, language-vi and version-vi in IAFA/Whois++ templates allow for group information related to a specific instance of the resource, tying them together with a common sequence number.

This standard provides encoding flexibility, not requiring a fixed set of elements. It introduces the cluster concept. A cluster is constituted by a set of data elements in order to group certain kinds of information such as name, address, telephone number, etc., that are often needed together. These data clusters may be associated with different roles of persons and organizations (author, administrator, owner, etc.) by a unique system identifier, enabling sharing and reuse of metadata elements. For efficient retrieval some of its elements follow a specific syntax, such as RFC 822 for electronic mail addresses and date/time.

3.3.2. Summary Object Interchange Format (SOIF)

SOIF was conceived as part of the Harvest architecture. Harvest is a distributed system for automatic resource indexing and discovery on the Web. It provides the ability to incorporate information in different formats from other sources, including high-quality, manually-created indexing information formats such as IAFA Templates and LSM (Linux Software Map) format. Its architecture is constituted by several subsystems. A gatherer subsystem collects indexing information and a broker subsystem (or Harvest indexes) provides a flexible interface for this information. Other subsystems enable facilities for indexing and searching resources that can be used by a variety of search engines [Bowman *et al.* 1994].

Records in SOIF format were designed to be generated by Harvest gatherers and used for user searches by Harvest brokers [SOIF]. The vast majority of SOIF templates in use today are automatically generated by robots acting as Harvest gatherers. SOIF is based on a simple attribute-value pair element, with a small number of common SOIF

```
@DELETE { }
@REFRESH { }
@UPDATE {
@FILE { ftp://www.des.ime.eb.br/~anamoura/hype96_7r_ps.gz
Title{84}: A Framework for Hyperdocument Generation in an Object Oriented
Database Environment
Author{34}: Moura, Ana Maria and Quadros, João
Time-to-live{7}: 9676800
Last-Modification-Time{9}: 774988159
Refresh-Rate{7}: 2419200
Gatherer-Name{50}: Personal Publications
Gatherer-Host{21}: leblon.ime.eb.br
Gatherer-Version{3}: 1.0
Type{10}: Compressed
Update-Time{9}: 774988159
File-Size{6}: 164373
MD5{32}: 43193942d53f5a8e4a7b4bcff7a415
Embed<1>-Nested-Filename{13}: hype96_7r_ps
Embed<1>-Type{10}: PostScript
Embed<1>-File-Size{6}: 28233
Embed<1>-MD5{32}: 84c123582c3d0754a39a78a78a7e2fb6d23
Embed<1>-Keywords{52}: hypermedia/hyperdocument/object oriented database
}
```

Figure 5. An example of SOIF templates.

attributes. A single SOIF stream can contain multiple SOIF templates, each of which has an URL pointing to the referenced resource and a number of elements describing other metadata. Each element has an attribute name, the length of the value in brackets, a colon delimiter, and then the value itself. Databases of SOIF templates can be searched by CGI scripts, accessible via a WWW browser.

In the example in figure 5, the @UPDATE, @REFRESH and @DELETE commands are part of the broker's collector interface, which provides an additional command level on top of SOIF. The number inside the brackets is a byte-count for the corresponding field, containing arbitrary binary data, which make it possible to span multiple lines and allow for non-ASCII characters. The information concerning the unnesting operation over the fields are placed in the "Embed" field [SOIF].

Despite being automatically generated by Harvest gatherers, SOIF templates may also be easily created manually, because they have a simple, common element set. Besides allowing binary objects to be embedded in the template, as their value length is explicitly represented in each element, SOIF enables the actual object to be embedded within a template attribute.

3.3.3. Dublin Core

Dublin Core [Weibel *et al.* 1995] is the result of the first workshop on metadata, held in 1995 by the OCLC (Online Computer Library Center)/NCSA (National Center for Supercomputer Applications). The workshop's purpose was to develop a metadata record to describe network electronic information, without prescribing a record structure. An initial consensus was accepted by the participants, in which documents (DLOs: Documents Like Objects) would be the first main element to describe. This means that, although in the Internet environment a document may be composed of texts, images, audio, video, or even another hyperdocument, a document should be considered as a whole object. The main goal of the model is to identify and define a syntax independent set of metadata elements to define the resources on the Web, in such way that they could also be mapped into more complex and highly controlled systems such as USMARC. Consequently, a set of thirteen basic elements has been defined, concerning:

- **intrinsicality:** to describe the object properties,
- **extensibility:** to allow inclusion of extra descriptive material for site-specific purpose,
- **optionality:** to ensure that all elements are optional,
- **repeatability:** meaning that all elements in the Dublin Core are repeatable,
- **modifiability:** expressing that each element in the Dublin Core has a definition that is intended to be self-explanatory, in order to satisfy the needs of different communities.

These characteristics are important because they aim to encourage authors and publishers to provide metadata to be collected via automated resource discovery; they are also intended to stimulate the creation of network publishing tools containing a template for metadata elements, simplifying the task of creating metadata records. Dublin Core is also projected to become a standard for promoting exchange between user communities.

```
<HEAD>
<TITLE>Metadata Support to ....</TITLE>
<META NAME="DC.title"      CONTENT="(TYPE=short) Metadata Support to Internet Resources ...">
<LINK   REL=SCHEMA.dc  HREF="http://purl.org/metadata/dublin_core_elements #title">
<META NAME="DC.subject"  CONTENT="(TYPE=keyword) metadata, resource description, ....">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements #subject">
<META NAME="DC.autor"      CONTENT="(TYPE=name) Moura, A.M.C">
<LINK REL=SCHEMA.dc  HREF="http://purl.org/metadata/dublin_core_elements#author">
<META NAME="DC.date"      CONTENT="(TYPE=creation)(SCHEME=ISO31)1997-07-30">
<LINK REL=SCHEMA.dc  HREF="http://purl.org/metadata/dublin_core_elements#date">
<META NAME="DC.autor"     CONTENT="(TYPE=email)  anamoura@ime.eb.br">
<LINK REL=SCHEMA.dc  HREF="http://purl.org/metadata/dublin_core_elements#author">
</HEAD>
<BODY>
```

Figure 6. An example of Dublin Core encoding.

Standardization of the metadata element set has been a dynamic process ever since this workshop. Active promotion of its results has been carried out to establish liaison with formal associations stakeholders, such as GIS and library communities, publishers, document vendors, SGML vendors, etc., in order to work on the problem of text encoding. At the CNI (Coalition for Networked Information)/OCLC image-metadata workshop [Weibel *et al.* 1997a], the initial set of core elements was modified to include image requirements such as a field for rights management and control, and another field for content description.

Figure 6 presents an example of formulation using Dublin Core's HTML implementation proposed in [Weibel 1996]. Qualifiers are defined as an unordered set of unique attribute-value pairs that are attached to each element [Knight and Hamilton 1996a]. The SCHEMA qualifier provides a way to interpret the field's content value according to some coding system, improving retrieval by introducing a level of standardization to the Dublin Core model; the TYPE qualifier is a mechanism that allows a Dublin Core element to be associated with several attributes: an author may have a name, a telephone number, etc. The <LINK REL> tag associates the metadata element to the naming authority (online or off-line source) which defined the element (it may be considered as a registry mechanism for metadata elements). A more detailed list of Dublin Core qualifiers has been proposed by Knight and Hamilton [1996a].

The Dublin Core specification is still being defined, but the appropriate use of qualifiers has introduced some new issues because they give great flexibility, while causing problems for the search mechanisms [Weibel *et al.* 1997b]. A syntactic foundation for wWeb-based metadata, the Resource Description Framework [RDF] was proposed at the 5th Dublin Core Conference held in Helsinki [Weibel and Hakala 1998].

Metadata used by IAFA/Whois++ and SOIF for discovery services tend to be based on simple record structures, such as attribute-value pairs. They do not contain an elaborate internal structure and do not easily represent hierarchical or aggregated objects; nor do they express relationships between objects. IAFA templates are the most detailed (there are templates for different types of objects), although they can only be used manually. SOIF provides automatic indexation, and it is able to index information manually from IAFA templates. Dublin Core represents the evolution of its two antecedents. It is based on MARC to describe the essential resource elements on the Web and it is allied to extensible architectures (Warwick, RDF) to provide object description in different hierarchical levels on the Web.

3.4. Metadata standards for global information infrastructure

GILS is the representative standard of this category. Its main purpose is to provide a mechanism for locating useful information generated by many government agencies. Some of its characteristics are described below.

3.4.1. Government Information Locator Service (GILS)

GILS [Christian 1996] was created as an initiative of the US federal government to help people find information resources throughout its many agencies. GILS identifies and describes these resources, supplementing other government and commercial information-dissemination mechanisms. In a broader sense, GILS can be defined as a decentralized collection of locators and associated information services used by the public to find information, either directly or through intermediaries. Servers acting as GILS locators are also information resources and can themselves be described by a GILS locator record in other GILS locators. GILS defines around 70 registered attributes (called GILS core elements), including title, originator, date of publication, place of publication, language, abstract, controlled subject index, spatial domain, among many others [GILS]. Since it adopts the ANSI Z39.50 standard protocol [Z3950] to specify how electronic network searches can be expressed and how results are returned, another 100 registered attributes, inherited from this protocol, are available. An important aspect of this standard is to ensure interoperability on a semantic level with the many different GILS servers.

title:	Metadata Support to Internet Resources Description and Retrieval
originator:	Ana Maria Moura
ContactNetworkAddress:	anamoura@ime.eb.br
ContactTelephone:	55-21-295 3232
abstract:	Metadata, or "information that makes data useful", have been ....
contactOrganization:	Instituto Militar de Engenharia
uncontrolledTerm:	metadata, resource description
linkageType:	text/html
linkage:	http://www.des.ime.eb.br/~anamoura
LanguageOfResource:	English

Figure 7. An example of GILS encoding.

Formulation example using GILS is given in figure 7.

GILS offers a complex metadata format due partly to its strong influence from MARC and Z39.50 communities, leading to a broad constituency of uses. Due to its use of these standards, GILS takes advantage of existing networks and software to access a vast array of important resources, such as libraries, museums, and archives world wide. According to GILS Application Profile, USMARC record format may be used to provide physical transfer of GILS records. It contains also a number of element subsets for dealing with simple geospatial and temporal metadata.

3.5. Additional considerations

Although some of the standards presented were not originally conceived for the Web environment, they have been extended to contemplate the description of network electronic resources:

- **MARC:** despite being designed for detailed bibliographic information, it has greatly influenced the conception of other standards. Recently an extra field has been included to describe the electronic location and access mode of Web resources.
- **TEI:** despite not providing a specific field to describe the location and access mode of electronic resources on the Web, TEI may be used as independent headers (stored separately from the text to which they refer) to describe network resources which are not necessarily TEI encoded. They can be manipulated, searched, and retrieved by any software dealing with SGML records. However, there is no provision for TEI headers within Internet-search and retrieval protocols.
- **EAD:** a database of EAD-encoded finding aids may become available on the Web. It can be loaded on an Internet server and manipulated by employing a search engine and a user interface. It provides viewers and it allows the use of commercial SGML software (such as DynaWeb) associated with an indexing tool.
- **SOIF:** already in use on the Web, SOIF-format records are generated by Harvest gatherers and used for user searches by Harvest brokers. The Harvest distribution contains a number of gatherer programs that can generate SOIF summaries from plain text, SGML, PostScript, MIF, and RTF formats.
- **DUBLIN CORE:** some guidelines to extend its elements have been proposed to create a generic application integrating all kinds of information resources [RLG 1997]. The tendency is to associate Dublin Core with metadata architectures in order to become the Web standard.
- **GILS:** because GILS Application Profile (which represents a set of terms that specifies the behavior of server software in conversation with client software) has no user interface, access to GILS servers must be accomplished through gateways, clients, or agents. GILS services can support interoperable search of many different metadata structures: HTML, SGML, X.500, SQL databases, PURL's (Persistent Uniform Resource Locator) [PURL], handles, Dublin Core, SOIF's, IAFA, Internet mail, DIF's (Directory Interchange Format) [DIF 1996], Whois++ templates, spatial metadata, etc. Whenever appropriate, servers simply map local semantics to the registered elements.

Some standards have more chance of being widely deployed on the Web due to their association with search and retrieval protocols, such as Z39.50, which already include search terms from MARC, GILS and Dublin Core. The success of a standard is directly associated with its ability to be flexible and adaptative in order to conform to user requirements. In fact, it is very difficult for a unique standard to provide all the requirements of various communities. Although some (such as Dublin Core) go in that direction, there is still a lot of work to be done in this field.

4. Using metadata for describing information resources

There is a general consensus within the information system community that the use of metadata is the main factor to promote integration and information exchange among the Web's heterogeneous resources. This integration requires a mechanism to represent semantic information supporting correlation of heterogeneous types of information.

Some metadata classifications have been presented in the literature [Bohm and Rakow 1994; Kashyap *et al.* 1995], either based on how metadata are obtained or on descriptors' functional characteristics. These classifications are important for understanding and defining the nature and scope of metadata. We first present a metadata classification based on the way metadata are obtained, followed by a discussion of functional classifications. This is an important step to present a more general classification integrating fundamental aspects and issues covered by other functional classifications. We believe classifications based on metadata use are more intuitive and spread elements more evenly through categories.

4.1. General metadata classification

This classification has been proposed by Kashyap *et al.* [1995] and it is based on metadata required by researchers when accessing different digital media types. It focuses on the nature of metadata, i.e., the way they are obtained, considering whether they are based on the document data content or not. The nature of the metadata is related to their content, media-type, and domain dependence. Kashyap identifies three kinds of metadata, well summarized in [Prabhakaran 1997]:

- **Content-dependent metadata:** these metadata depend only on the content of media objects, i.e., they are automatically extracted from the object content. Derivation of facial features of a person's photographic image (such as type of nose or ear, color of hair) and derivation of camera operations (such as panning, tilting and zooming) in a video clip belong to this category.
- **Content-descriptive metadata:** these metadata associate descriptive terms with the resource content, but cannot be generated automatically from this content alone. This type of metadata describes the characteristics of media objects based on impressions created by the user or an application. For example, metadata on facial expression such as anger or happiness derive from the content of the image, but they depend on the user for their definition or either on tools which can support such cognitive process.
- **Content-independent metadata:** these do not depend on media information content, but are associated with it. A photographer's name, last modification time, location of a document, or a movie budget are all examples of this type of metadata.

The definitions above sometimes cause some misunderstandings since content descriptive metadata (which give an interpretation for a resource content) could also be considered dependent on content.

4.2. Functional classifications

Papers related to metadata commonly use a functional classification to identify metadata components. Bearman [1996] presents a functional classification as a reference model to enable electronic business interaction. In the review of metadata standards made by Heery [1996a], this kind of classification is used to identify the main goal of each one of the standards. The FGDC standard [FGDC] also organizes its components in sections based on a functional classification.

4.2.1. A metadata model for describing multimedia documents

An electronic document may be composed of different media types (such as video, text, image, sound), each one playing a specific role in the document context. When considered as individual components, they present their own semantic features. Answering a user query, for example, typically requires correlation of relevant information across multiple forms and representations, which may even be stored in different repositories. The approach proposed in [Kashyap *et al.* 1995] uses metadata to establish the correlation, at a higher semantic level, amongst heterogeneous types of information. This model enables the representation of specific features according to their media type, starting from basic concepts.

The classification presented in this section focuses on the use of metadata for multimedia items [Bohm and Rakow 1994]. Six functional categories are identified:

- **Metadata for representing media types:** they provide information in connection with the presentation of multimedia data, language, format, coding and compression techniques.
- **Content-descriptive metadata:** this metadata are used to describe the content of a multimedia document and they can be media-type dependent.
- **Metadata for content classification:** they correspond to additional information that can be derived from the document content. In medical context, these information can describe, for example, the level of expertise required from the reader in that domain.
- **Document composition metadata**: these metadata describe knowledge about the relationships between document components and the role of each component in the document.
- **Metadata for document history:** they provide information which is related to the history and the status of the document, concerning its individual components.
- **Metadata for document location:** this information allows access to multimedia data.

When compared to the previous classification, the categories 1, 5 and 6 can be identified as content independent metadata, while the third one is considered domain dependent content-descriptive metadata.

4.3. Functional classification based on metadata component types

In this section, we present a classification of metadata elements which tries to integrate some of the approaches mentioned above. This classification will be used in section 4.4 for a comparison of existing metadata standards.

4.3.1. Metadata for resource discovery

These metadata represent the set of terms necessary to discover and identify a resource on the Web [MARBI 1995]. This is the main purpose of some models such as Dublin Core, which describes the fundamental metadata elements to identify a resource, including title, author, identifier, and subject. In this model, some additional bibliographic information is included to ensure that the resource in question is actually the desired one: contributors, related dates, and resource type and category (such as a novel, a technical report or a poem). Also, technical characteristics, such as language and format (PostScript, HTML, etc.), could be used to select a specific instance of the resource. Relation and source elements allow the association of the resource with others, like its bibliographic references and original sources. The USMARC format for bibliographic data provides a complete set of metadata elements to distinguish one resource from another. Unlike the Dublin Core Model, it requires some specialized knowledge due to the complexity of its use. Generally, metadata for resource discovery include:

- **Basic bibliographic descriptive metadata:** although varying according to the model intended purpose, most models present basic descriptive components for identification (like title), for responsibility (such as authors and contributors) and extra information to better identify and characterize the resource (edition and series information, for example).
- **Metadata for resource unique identification on the Web:** these metadata elements correspond to an extension of the basic bibliographic descriptive metadata adapted to the Web. They provide a unique identification for Web resources and they will be supplied by a naming schema authority and solved by a resolution system. Some of those schemas include: RCDS (Resource Cataloguing and Distribution Service); the Handle System; X-DNS-2 (based on the Internet Domain Name System); URN services; URN path (also making use of the DNS); and Whois++ [URN 1996]. Other globally unique identifiers, such as ISBN (International Standard Book Number) or ISSN (International Standard Serial Number), may also be referenced in this element. Dublin Core model considers the resource version number as an additional identification element.
- **Metadata for general content description:** these metadata correspond to elements describing the resource contents. Dublin Core and IAFA/Whois++ models include the field description for this purpose. A simple example is an abstract of the resource in descriptive prose for textual media. This field may also be used to reference another object containing a more complex descriptive schema or the description under a different media (a video exhibition containing explicative information, for example).
- **Metadata for subject description:** these elements may include controlled or uncontrolled terms, used to characterize the information contents of the resource. Usually they correspond to a set of keywords or a subject descriptor. For spatial and temporal based resources such as Cartographic or GIS sources, these metadata also include spatial and temporal coverage information (latitude and longitude ranges, for example).
- **Metadata for structure description:** these metadata are used to describe document composition (images, chapters in textual documents, etc.). For textual media, a table of contents may be used to describe the document structure. Ferber [1997] discusses a metadata model for hyperdocuments considering typed links. He also considers the recursive structure of the Warwick framework to represent a hypermedia structure for a WWW document.
- **Metadata for relationship description:** these metadata include elements to associate a specified resource with other related documents. Dublin Core includes the relation element which may be used for this purpose.
- **Metadata for describing resource provenance:** these metadata identify the primary sources or providers of data to the system. Dublin Core model uses this element to express special relationships with others having the same intellectual content.
- **Metadata for format and media description:** these elements include presentation characteristics and data representation of the resource components: media type, size, format, compression standards, etc. They may include media specific information such as the number of channels and the sample rate for audio or the number of frames and shots for video. Bohm and Rakow [1994] characterized the language attribute as specific for textual media. The Dublin Core model associates the format field with an Internet Media Type (or MIME content type) assuming values like text/sgml, text/html or image/giff.

4.3.2. Metadata for resource availability

These metadata define the terms and conditions required to access and to retrieve the resource, no matter if in a restricted or unrestricted way. As stated in [Bearman 1996], these metadata may be textually described or they may specify the unique identification(s) of resolvers containing the terms and conditions for accessing and using the resource.

- **Metadata for resource distribution:** these elements describe how the information resource is made available and may specify its medium, distributor(s), publication details, contact, order process information (including payment requirements), among others. The distribution section in FGDC, for example, specifies fee information and how to obtain the resource in a non-digital or digital form (online or off-line options).

- **Terms and conditions for resource access:** these elements specify if there exist previous access conditions to be fulfilled to assure the protection of privacy and intellectual property, such as access authority permission, and system identification requirements (logon, password, etc.).
- **Terms and conditions for resource use:** these metadata describe the terms and conditions for using the resource such as a copyright notice, restrictions or policies for copying, modifying, and others. They may define or reference specific views of the resource according to user access permissions.
- **Metadata on resource requirements:** these metadata specify the software and hardware conditions for resource use, such as special viewers, environment settings and configurations.
- **Metadata for resource location:** these metadata provide the necessary information for transferring the resource to a local site, such as information about where to locate a specific manifestation of the resource, including its URL and access protocol. The main purpose is to allow the system to select an appropriate copy or version of the resource, reducing network overload by considering cost and location aspects, for example.
- **Metadata on resource authenticity:** these metadata describe schemes or methods to ensure the authenticity of a resource. They include both simple mechanisms (based on file size) and sophisticated ones (based on digital signature). For this purpose, SOIF templates specify the MD5 element.

4.3.3. Metadata for resource usage

These are additional information to allow the resource adequate usage.

- **Metadata for resource content classification:** these metadata are based on resource content to classify a resource according to some contextual schema assigned by a rating authority. PICS platform [PICS 1996], for example, provides a way to associate rating labels with a document content. Some standards, like GILS, include a security classification control associated with the information resource (top secret, confidential, etc.).
- **Metadata for describing resource data quality:** these metadata specify the quality of the data, which is especially important to some areas like GIS applications. FGDC standard, for example, provides a section concerned with this category of metadata. These metadata may include any information relative to the validity, degree of reliability and error estimate of a specific resource. Information about the resource lineage may also be specified, as observed in SAIF standard.
- **Metadata for describing resource purpose:** these metadata describe why the information resource is offered, identifying programs, projects, discussion forums, etc., related to this resource.
- **Metadata for resource contextual description:** these metadata provide information related to specific events, situations, settings, etc., related to the resource domain or purpose. TEI standard, for example, includes elements for contextual information: <textDesc> gives a complete description of the situation in which the text was produced, such as domain, revision, derivation, factuality; <partDesc> describes the identifiable speakers, voices or other participants in a linguistic interaction; and <settingDesc> describes the setting(s) within which a language interaction takes place [Burnard 1994].

4.3.4. Metadata for resource administration and control

These metadata provide the information to control, audit and trail the information about the resource itself as well as its metadata.

- **Metadata for resource modification control:** these are metadata for version control and may specify: modification and review dates, modifications introduced, contact information of the modifications' authors, etc. In the TEI standard, the header provides the <revisionDesc> element to record a detailed change log or a revision history of the resource.
- **Metadata for resource administration:** these metadata are related to any information concerning the management and control of the resource itself (creation date, valid from and valid to dates, resource administrators, contact information, etc.).
- **Metadata for resource use history:** these metadata are reserved to store information about the operations performed on the resource such as copy, edition, removal, etc. They should also specify the operation executor(s), as well as the operation date and time [Bearman 1996].
- **Metadata for metadata administration:** these metadata are related to any information concerning the management and control of the metadata itself (creation date, last review date, next review date, record administrators, contact information, language, standard name, standard version, etc.).

4.4. Metadata components according to the functional classification

Using the metadata functional classification presented above, it is possible to compare the standards and models described in section 3 according to their use and element set coverage.

Figure 8 contains a summary overview of this comparison. As figure 8 shows, there are many types of metadata elements common to all standards discussed. Most of them offer a very comprehensive cover of the functional sets, although some present only few elements in each set.

FUNCTIONAL SET		IAFA/ Whois++	USMARC	Dublin Core	TEI	SOIF	EAD	GILS
Metadata for Resource Discovery	Basic Bibliographic Descriptive Metadata	x	x	x	x	x	x	x
	Metadata for Resource Unique Identification	x	x	x	x	x	x	x
	Metadata for General Content Description	x	x	x		x	x	x
	Metadata for Subject Description	x	x	x	x		x	x
	Metadata for Structure Description		x				x	x
	Metadata for Relationship Description		x	x	x	x	x	x
	Metadata for Describing Resource Provenance	x	x	x	x	x	x	x
	Metadata for Format and Media Description	x	x	x	x	x	x	x
Metadata for Resource Availability	Metadata for Resource Distribution	x	x	x	x		x	x
	Terms and Conditions for Resource Access		x				x	x
	Terms and Conditions for Resource Use	x	x	x		x		x
	Metadata for Resource Content Classification		x					x
	Metadata on Resource Requirements	x	x			x		x
	Metadata for Resource Location	x	x	x		x		x
	Metadata on Resource Authenticity					x		
Metadata for Resource Usage	Metadata for Describing Resource Data Quality		x					x
	Metadata for Describing Resource Purpose		x					x
	Metadata for Resource Contextual Description				x		x	
Metadata for Resource Administration and Control	Metadata for Resource Modification Control	x		x	x	x	x	
	Metadata for Resource Administration	x		x			x	
	Metadata for Resource Use History							
	Metadata for Metadata Administration	x	x			x		x

Figure 8. Standards and their functional metadata set.

5. Retrieving resources on the Web

A vast number of Internet retrieval services have emerged in the last five years. The best known services (such as Lycos, Excite, Alta Vista) are generally performed by so-called robots. They aim to be global indexes, covering all countries, subject areas, and a number of different information protocols. Some aspects of these global, generally robot-based services have evolved so fast that they require great agility by the agencies to keep repositories updated, in order to satisfy their users' real needs. Here we refer to features such as quantity of indexed documents; geographical and subject area covered, as well as the number of facilities (update and retrieval capabilities) provided.

Koch *et al.* [1996] have defined the navigation process in three ways: *surfing*, which requires unsystematic following of links; *browsing* as the systematic "turning of pages," making use of structured information overviews and collections; and *searching* for the navigating course with the help of databases built for this purpose, allowing direct access to individual documents. In this section we present some of the main current Internet resource- retrieval services, utilizing Koch's terminology.

5.1. Browsing services

These services use intellectual classification and manual cataloguing in order to organize resource retrieval on the Web. In spite of their usefulness in certain cases, these services offer limited alternatives as they cover only a small fraction of available resources. Besides, due to the Internet's dynamic nature, catalogue maintenance becomes a very complex task. Some browsing services may be cited: Yahoo, as a subject browsing service; W3 servers as a geographical browsing service; and CUI World Wide Web Catalogue as a chronological browsing service.

Currently, many services are combining browsing structures with search capabilities, covering a more extensive area than their own browsing databases. Yahoo is an example of a very popular browsing service, which is now offering the Alta Vista index. On the other hand, search services like Lycos offer a complementary, subject-oriented browsing system [Koch *et al.* 1996].

5.2. Information gateways

An information gateway provides resource discovery based on subject information, which is filtered according to selective conditions, combining manual cataloguing with simple descriptive registers, like IAFA templates. Some projects follow in this direction: SOSIG (Social Science Information Gateway); OMNI (Organizing Medical Networked Information); EEVL (Edinburgh Engineering Virtual Library), ADAM (Art, Design Architecture and Media information gateway) and IHR-info (Institute of Historical Research) [Dempsey 1996].

5.3. Search services

Robots are systems created to collect and update resource indexes automatically from different servers connected to the Web. Robots have been classified according to their typologies described in [Koch *et al.* 1996]. Some of them index Web individual resources (persons, institutions, mail lists, software), or even other protocols, providing characteristics of different spaces on the Web (Gopher, Telnet, Wais, Z39.50, etc.). In addition to comprehensive services such as Alta Vista, Harvest, Excite, Hotbot, WWW Worm, and Inktomi, there are also regional services such as Nordic WWW Index, Finnish WWW Index Trampoliini, and JORI. There are even others based on subject, such as Europe Physics Broker and the robot-generated index of Nordic Libraries' WWW pages.

These services may be classified basically into three groups, according to indexed information:

- Indexing some document components, such as title and subtitles, or the most heavily weighted words in the text (selected according to an algorithm that determines their frequency or placement), among others. This is the case of Magellan Internet Guide and WWW Worm, Inktomi, among others.
- Indexing the full text content and providing a high level index, such as Harvest, Alta Vista, Excite, Open Text Web Index, HotBot, WebCrawler, InfoSeek Guide, and Lycos.
- Indexing meta information: these tools capture the information included in META-tag attributes by the document's author (or authors), as discussed later in section 5.6.

5.4. Search engines based on templates or lists

These generated indexes are based on templates and lists such as ALIWEB and Intercat.

ALIWEB [ALIWEB] is a system that automatically combines distributed WWW server descriptions into a single database. So basically it does for the WWW what Veronica does for Gopher, or Archie does for anonymous FTP. ALIWEB is based on the IAFA standard, and uses the site.idx file for registering URLs, etc. In this file, tags identify the type of the record.

5.5. Architectures for distributed searching

These architectures allow distributed indexing services among servers, enabling more structured information on the Internet. Whois++ [Faltstrom 1998; Roszkowsky and Lukas 1998] follows this category. It is a distributed directory service defined by the IETF group. This approach defines a mechanism to collect information in several servers, passing them again, in a collaborative way, to different index servers on the Web. CIP (Common Indexing Protocol) protocol, based on Z39.50, is used for communications between servers. A metadata architecture definition on the Web would be a basis for implementing specific directory services (such as subject, author, etc.).

5.6. Metadata-based services

Given the exponential increase of Web resources, search mechanisms require more effective solutions: metadata on the Internet should be available to allow search services to provide better and more precise retrieval results. HTML language already supports a limited metadata resource definition. META tags are placed in the head of the HTML document, between the <HEAD> and </HEAD> tags. Such information can be extracted by servers/clients for use in identifying, indexing, and cataloguing specialized document meta information. This is especially important when the document uses frames or when it has graphical components with no text for search engines to index (except that some may index the text inside the ALT attributes in the IMG tag, that connects an image to an HTML page [Vancouver Webpages 1997]).

The META element is used as a metadata container, describing document properties in attribute-value pairs. META tags have two attributes:

- <META NAME="name" CONTENT="content">,
- <META HTTP-EQUIV="name" CONTENT="content">.

The NAME attribute is used to name a property such as author or publication date and CONTENT specifies a value for the named property.

Example. < META NAME= "author" CONTENT= "Barreto, C.M.">.

The HTTP-EQUIV attribute binds the element to an HTTP response header. It can be used in place of the NAME attribute and has a special significance when documents are retrieved via HTTP. Some HTTP headers include: expires (to determine when to fetch a fresh copy of the associated document); refresh (specifies a delay in seconds before the browser automatically reloads the document); content-language (native language of the document); PICS-label (used to declare a document's rating in terms of adult content – sex, violence, etc.). HTTP headers should not be created without checking for conflict with existing ones since it is possible to interfere with a server operation.

Example. <META HTTP-EQUIV= "expires" CONTENT= "SAT 15 May 1998 09:13:25 GMT"> will result in the HTTP header: Expires: SAT 15 May 1998 09:13:25 GMT.

Some of the current WWW indexing services (in particular, AltaVista and Infoseek) are already supporting a limited metadata set (two elements: Description and Keywords). When these services index a WWW site, they look for these META tags; and they index the document based on an author-supplied list of keywords (they also index the rest of the document, but the keyword list takes precedence). The Description field is normally returned as part of the result-set [Iannella and Waugh 1997]. Examples of META tags for these two services are shown below:

- <META NAME="description" CONTENT="Metadata Support to...">,
- <META NAME="keyword" CONTENT="metadata, resource description">.

The Metadata Search Engine allows searching for metadata attributes on WWW pages. The Metadata Search Engine is a modification of the Harvest information discovery and access system. Whereas Harvest summarizes the contents of a document and matches queries to these summaries, the Metadata Search Engine extracts only predefined metadata marked-up using the HTML tag from the head of HTML pages. Queries are then matched with the metadata, not with a general summary of the document [MSE]. The Metadata Search Engine currently recognizes the HTML markups of a number of different metadata standards including Dublin Core, Alta Vista and Erin.

When metadata become more common (either embedded in documents, such as the META tag in HTML files, or from a separate metadata repository) and when indexing services start to concentrate on indexing this information, we should see a marked increase in the effectiveness of information retrieval. Author-generated (or even semi-automated) metadata will lead to increased quality because this information will certainly become more reliable and rich than automatically generated output.

Another advantage is that a user will then be able to do fielded searches. For example, a search like "Author=Smith and Subject=Metadata" would dramatically improve search results, as it would ignore the documents where these two words appear just as free-text.

The above initiatives represent the beginning of new possibilities of having metadata-aware search services. In reality, we might see in the near future a mix of full-text indexing and metadata-based content search as a good approach to improve resource retrieval on the Web. Developing this kind of service would require thorough understanding of the metadata role on resource description. The effectiveness of such services would largely depend upon the "good will" of information providers in having descriptors associated with their resources. For this, metadata should be made available according to well known standards.

6. Other metadata initiatives

As described in section 4, metadata are classified according to their purpose and to the resource they describe, which may vary enormously from one resource to another, making this task very difficult. The solution seems to be to construct a general architecture able to implement any metadata element set which supports administration and access to Web resources. Some initiatives in this direction have already started, as described below.

6.1. Metadata frameworks

There is a considerable number of more comprehensive initiatives based on the standards described above. Many of them consider a particular metadata standard, either because of its simplicity or extensibility, or even because some of them address one particular subject area.

6.1.1. STARTS

STARTS stands for Stanford Protocol Proposal for Internet Retrieval and Search [Gravano *et al.* 1997]. It comprises an architecture to support metasearchers (unified query interfaces to multiple search engines). Based on the SOIF standard, it defines the metadata that each source should export to describe its contents and capabilities.

6.1.2. ROADS

The United Kingdom created the Electronic Libraries Programme (eLib) as a three-year initiative to modernize library services at the higher education level. ROADS (Resource Organization And Discovery in Subject-based services) – is a development project in the area of Access to Network Resources of eLib. The objective of the project is to "investigate the creation, collection, and distribution of resource descriptions to provide a transparent means of searching for and using resources" [Heery 1996b]. ROADS is in the ambit of DESIRE [DESIRE], a project within the Telematics for Research area of the European Union's Fourth Framework Programme. DESIRE addresses the needs of European researchers to locate and retrieve information. It includes both the development of a robot-based Web index and the construction of subject-based resource catalogues.

ROADS uses IAFA templates, but has a component to provide interoperability with MARC and Z39.50 protocol. ROADS tools include a search engine, several tools for creation and processing of IAFA records, a Web interface, and a mechanism for selecting between servers to search.

6.1.3. NORDIC metadata project

This project involves many Scandinavian institutions with the objective of creating a metadata production, indexing and retrieval environment. Some of the project's tasks include: enhancement of the Dublin Core specification (mostly by defining new schemes for some of the elements), the conversion of Dublin Core to Nordic MARC formats and vice versa, the creation of recommendations for Dublin Core syntax issues on HTML, the definition of some basic requirements for user interfaces, and the development of metadata-aware search services (possibly adapting the already-in-use Nordic Web Index) [Nordic 1996].

6.1.4. National Spatial Data Infrastructure

The National Spatial Data Infrastructure (NSDI) is a part of the United States National Information Infrastructure Initiative, especially focused on georeferenced information. The NSDI developed a set of guidelines for implementing metadata and data services in this area. These included using the FGDC proposed standard as the basis for all infrastructure efforts. Also, the NSDI emphasized the need to decentralize metadata and data holdings to assure better quality and currency.

Under this initiative, several government agencies are providing digital spatial data together with standardized metadata on the Internet. A related project is the GeoWeb project from the State University of New York at Buffalo, providing a gateway to discover and access distributed spatial data using a common set of tools. GeoWeb has been developing a central metadata repository using the FGDC standard. Another related project is the Alexandria Project described below.

6.1.5. The Alexandria Project

The Alexandria Project [Goodchild 1996], originally established in 1994 at the University of California, Santa Barbara, involves several US universities, libraries, and agencies, with the primary goal of designing, implementing, and deploying a digital library for spatially indexed information. The project includes a metadata strategy based on a hybrid of FGDC and USMARC standards. It uses FGDC as a baseline and supports selected USMARC fields (basically those for maps and computer files) as extensions. Research in this project includes innovative methods of data storage and compression, geographically indexing the stored information, and providing a comprehensive user interface for utilizing map and image resources. A first prototype, called ARP (Alexandria Rapid Prototype) was built to allow evaluation of the main issues underlying digital libraries for spatial information.

6.1.6. The Dienst Project

The Dienst Project is a conceptual architecture for digital libraries, a protocol for communication in that architecture, and a software system implementing that protocol to provide Internet access to distributed, decentralized document collections in multiple formats [Lagoze and Davis 1995]. Document descriptions basically concerning bibliographic information are stored in RFC 1807 format [Lasher and Chen 1995]. Dienst architecture contains four classes of services, all of them communicating via Dienst protocol: a repository service to store digital documents; an index server to search a document collection; a centralized meta-service that provides a directory location of all services; and a user interface to mediate user access to this library. The NCSTRL (Networked Computer Science Technical Report Library) collection is currently the most significant collection using Dienst protocol.

6.2. Other initiatives involving several standards

Many projects under way admit that it is unlikely that some unique metadata format will be universally used. They recognize the need for a higher-level container architecture that can accommodate different metadata standards already in use. Besides, most of them allow mapping and conversion between standards. Some other proposals address more basic features like naming and organizing digital objects; and establishing general frameworks where many different initiatives could be accommodated. This is the case of the InfoHarness Repository Definition Language [Shklar *et al.* 1995] and with the Handle Server Infrastructure proposed in [Kahn and Wilensky 1995].

Among the projects on higher-level container architectures, a common approach is to identify the description scheme in some way, together with the description. To ensure that the architecture supports new user services, metadata representations usually consist of an extensible set of specific information service components. Generally, the system is flexible enough to understand new schema and to follow links from documents to external metadata.

Here follows an examination of some initiatives based on these ideas.

6.2.1. URC (Uniform Resource Characteristics)

URCs are descriptions of Internet-accessible resources. Initially, the IETF working group developed the concepts of URCs, associating a URN (Uniform Resource Name) with a set of URLs used to fetch the resource [URC]. As the URN service has not become available and in order to accommodate different communities' requirements, the URC proposal has evolved to allow different sets of resource descriptions, called URC *subtypes*. Very few useful elements (based on the Dublin Core Proposal) are mandatory such as URL, location, URN, identifier, instance, and format. A canonical URC representation and basic operations to manipulate it are currently being defined in order to accommodate several formats that can be mapped into and out of a variety of syntax forms. A transformation language would provide an appropriate translation from one subtype to another. URC is now an abstract structure: it is represented by a decorated tree with typed nodes. Each node may have an identification and associated attribute/value pairs (title and author, for example, being the tree decorations). On-node operations, such as tree construction and destruction, and tree search operations, are also being defined [Daniel 1996a].

The URC proposal is intended to be mapped to popular protocols, such as HTTP and Z39.50, and may specify how operations on the abstract model can be encoded into different standard transfer protocols, assuring a desired level of interoperability. Four levels of conformance are being defined to implement the capabilities of the full canonical representation [Daniel 1996b].

6.2.2. Warwick Framework

The Dublin Core working group recognized that different metadata definitions would be in use on the Internet. They suggested an infrastructure conceived to support any metadata element set. This infrastructure is called the Warwick Framework and it is based on a container architecture to aggregate distinct packages of metadata [Lagoze *et al.* 1996]. The architecture supports autonomous administration and access to metadata packages.

The basic components of the Warwick Framework, represented in figure 9, are [Lagoze *et al.* 1996]:

- **Container:** the basic unit where sets of typed metadata (the packages) are aggregated.
- **Package:** the set of typed metadata representing a first class object. The package content is simply considered a sequence of bits. They can be divided in three categories:
  - **Metadata set** – a separately defined primitive metadata format. They contain actual metadata. These can be MARC records, Dublin Core records, and/or others.
  - **Indirect** – a reference to an external object. The target of the reference is a first-class object, and thus may have its own metadata and conditions for access. Indirect packages allow the sharing of metadata objects, as the target of the indirect package may also be indirectly referenced by other containers.
  - **Container** – a package that is itself a container, with no defined limit for this recursion.

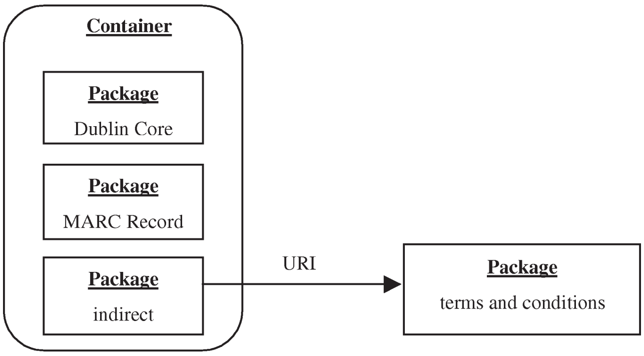

Figure 9. Example of a Warwick Framework contained [Lagoze *et al.* 1996].

An extension of this framework has been proposed in [Daniel and Lagoze 1997]. Its main concept is based on the use of distributed active relationships (DARs). These relationships provide a model for representing data and metadata in digital library objects, without any essential distinction. DARs leverage the connectivity and computational characteristics of networked environments to create dynamic relationships between data resources.

Implementations of the Warwick Framework were proposed in HTML [Miller 1996a,b], SGML [Burnard *et al.* 1996], MIME [Knight 1996b] and, more generally, as distributed objects. This last one allows for the implementation of the full capabilities of the Warwick Framework. It is based on the requirements specified by Kahn and Wilensky [1995] for a digital library architecture. More recently, a digital object and repository architecture, named FEDORA [Payette and Lagoze 1998] was implemented. One of its main features is to support heterogeneous data types. Its theoretical foundations lie in the Kahn and Wilensky and Warwick frameworks.

6.2.3. PICS (Platform for Internet Content Selection)

PICS was initially conceived to solve the problem of child protection on the Internet by providing a common format for labels, so that any PICS-compliant selection software can process any PICS-compliant label [PICS 1996]. This initial idea has evolved to address general network metadata requirements. Currently, PICS can be considered as a concrete framework for transporting different sets of metadata related to Internet resources. PICS provides a method for naming (via URL) and describing a metadata system (e.g., Dublin Core model) as defined by Rating Services and Rating Systems; and a method for encoding metadata as specified in PICS Label Distribution Label Syntax and Communication Protocols. Metadata labels may be distributed inside an HTML document; with a document transported via any protocol that uses RFC-822-style headers; or even separately from the document.

For transporting generic metadata, an extension for PICS became necessary to solve its inability to handle text-based metadata within labels. Support was required for: string data type, repeated values, structure, extensibility, and grouping and preserving the ordering of metadata elements. To address these problems PICS-SE, considerably influenced by object-oriented technology, provides a way to transport classes and annotation objects as PICS labels, not requiring any change in current PICS syntax [Braum *et al.* 1997]. More recently W3C has been committed to provide reference code for important components of this architecture, including among others, resources for developers of software and labeling services, such as a standard library in JAVA. This library parses PICS labels, service descriptions, and profiles [PICS 1998].

6.2.4. Stanford Digital Library

The Stanford Digital Libraries project is part of the Digital Library Initiative, started in 1994 and supported by the NSF, DARPA, and NASA. It is a big project, including five other universities besides Stanford and a number of industrial partners. At the heart of the project is the InfoBus protocol, which provides a uniform way to access a variety of services and information sources through *proxies* acting as interpreters between InfoBus and native protocols. InfoBus consists of distributed objects that communicate with each other through remote message calls, using CORBA specifications. A variety of user level applications provide powerful ways to find information, using either cutting-edge user interfaces for direct manipulation or Agent technology.

To facilitate metadata compatibility and interoperability in such a digital library, InfoBus relies on a metadata architecture that includes four basic component classes: attribute-model proxies, attribute-model translators, metadata facilities for search proxies, and metadata repositories [Baldonado *et al.* 1997]. *Attribute-model proxies* represent a real world attribute model, with a set of attributes corresponding to an existing metadata standard. They encapsulate information that is specific to an attribute model. *Attribute-model translation services* serve to mediate among different metadata conventions that are represented by attribute-model proxies. These translation services know how to translate attributes from one model to another. The *metadata information facility* refers to each search-service proxy supported and it is responsible for exporting metadata about the proxy as a whole, as well as for exporting metadata about the collections to which it provides access. It includes information such as the collection name, a link (URL) to a content summary of the collections, the attribute models supported, details about the search service (such as Boolean operators supported), and access constraints and descriptors relative to the metadata generation itself. All metadata available from attribute-model proxies, attribute-model translation services, and search-service proxies are collected in one central database – the *metadata repository.*

6.2.5. Meta Content Framework (MCF)

The Meta Content Framework [Guha 1997] is a structure-description language based on an extensible data model presented by Netscape [Netscape] to the World Wide Web Consortium (W3C). It was conceived to be used for describing the structure of Web sites and any information source on these sites. MCF is a comprehensive proposal, that can serve multiple purposes:

- its rich information can be used to support more powerful search engines, allowing concept-based searches,
- a robot could use it to determine which portions of the site to index,
- a browser could use it to present a site map,
- a channel client could use it to periodically download portions of the site.

The MCF data model includes a set of basic types that can be extended to accommodate new kinds of data or different application needs. MCF also defines a basic vocabulary that includes a set of terms commonly used for describing the content of Web documents. These terms were largely derived from existing standards such as the Dublin Core standard.

MCF is represented using an XML based syntax. XML is an extensible, general purpose, data-representation language. XML Hyperlink is used to refer to externally stored MCF blocks. These are important as they allow sharing and re-use of descriptions, avoiding duplication. For HTML pages, the HTML Link element can be used to associate MCF files.

6.2.6. Resource Description Framework (RDF)

RDF [RDF] is a W3C initiative to provide a unifying architecture for processing metadata and interoperability between applications that exchange information on the Web. It accommodates the diversity of semantics and structure needed by various communities. For example, in resource discovery to provide better search-engine capabilities; in cataloguing for describing content and content relationships available on a particular Web site, page, or digital library; by intelligent software agents to facilitate knowledge sharing and exchange; and for describing Web-page intellectual property rights, among others. It has been substantially influenced by the Warwick Framework and uses XML as the encoding syntax for metadata.

7. Conclusion

Adequate metadata management is an essential requirement to provide effective use of information resources in the ever-expanding, heterogeneous, distributed Web environment. This paper surveyed many proposals presented in recent years to try to establish some common ground for the different description standards already in use and associated with network resources. A detailed functional classification of metadata components has also been presented as a basis to compare existing proposals.

The different initiatives focusing on the use of metadata on the Web have revealed some general tendencies as well as a proliferation of metadata standards, serving different needs. Dublin Core certainly served as the basic structure upon which many other proposals have been developed. Since the Fourth Dublin Core Metadata Workshop [Weibel *et al.* 1997b] there has been a consensus about some modifications of element names and definitions and, more recently, about its integration to the RDF framework [Weibel and Hakala 1998]. With these, Dublin Core would adequately serve for the description of a large class of resources, particularly those sharing characteristics with document-like objects, Dublin Core's original focus. Also, more complex descriptions usually consist of extensions to Dublin Core or at least provide some kind of mapping to its elements. Conversion mechanisms from one standard to another seem to be a necessity given that different standards are bound to coexist.

Another aspect made quite clear in this study concerns the evolution of Web metadata architecture to support the encoding and transportation of many independently developed varieties of metadata to maximize system interoperability. There are currently several initiatives underway, under the auspices of the W3C, to address these issues:

- HTML 4.0 [HTML] extends the previous version increasing its potential with the addition of new features.
- The Extensible Markup Language (XML) gains more and more users every day. XML retains the key SGML advantages of extensibility, and structure and validation, but it is designed to be vastly easier to learn, use, and implement. It also provides the additional capability of incorporating Java-based Web applications to XML.
- PICS-SE, which stems from the same motivation as the Warwick Framework, is ready to use: it provides a concrete syntax as well as a mechanism to transport metainformation [PICS 1998].

Most of all, one can predict significant improvements in the network resource discovery domain, like static surrogates replacing dynamic ones to support search processes [Lagoze 1997].

References

AACR (1988), "Anglo-American Cataloguing Rules," Second Edition, Revision.

ALIWEB, `http://www.nexor.co.uk/public/aliweb/.`

Anderson, J.T. and M. Stonebraker (1994), "Sequoia 2000 Metadata Schema for Satellite Images," *Sigmod Record 23*, 4, 42–48.

Baldonado, M., C.C. Chang, L. Gravano, and A. Paepcke (1997), "The Stanford Digital Library Metadata Architecture," *International Journal of Digital Libraries, 1*, 2.

Barreto, C.M. (1998), "A Metadata Model for Describing Electronic Documents on the Web" (in Portuguese), MS thesis (in progress), Department of Computer Engineering, IME, Rio de Janeiro, Brazil.

Bearman, D. (1996), "Metadata Requirements for Evidence."
`http://www.sis.pitt.edu/∼nhprc/.`

Beckett, D.J. (1995), "IAFA Templates in use as Internet Metadata."
`http://www.hensa.ac.uk/tools/www/iafatools/paper/paper.html.`

Bohm, K. and T.C. Rakow (1994), "Metadata for Multimedia Documents," *Sigmod Record 23*, 4, 21–26.

Bowman, C.M., P.B. Danzig, D.R. Hardy, U. Manber U., and M.F. Schwartz (1994), "Scalable Internet Resource Discovery: Research Problems and Approaches," *Communications of the ACM 37*, 8, 98–107.

Braum, I., A. Konig, and T. Wichmann (1997), "PICS-SE: A Proposed Standard for Annotation of Internet Documents using a String Extension to PICS."
`http://www.kulturbox.de/aid/pics-se/.`

Burnard, L. (1994),"The Text Encoding Initiative Guidelines."
`http://www.uic.edu/orgs/tei.`

Burnard, L., E. Miller, L. Quin, and C.M. Sperberg-Mc Queen (1996), "A Sintax for Dublin Core Metadata: Recommendations from the Second Metadata Workshop."
`http://users.ox.ac.uk/∼lou/wip/metadata.syntax.html.`

Christian, E.J. (1996), "GILS What is it? Where's it going?," *D-Lib Magazine*, January.
`http://www.dlib.org.`

CIP, "DPRS Technical Note/002."
`ftp://styx.esrin.esa.it/pub/od/CIP/other/cipz39.tn.ps.`

Daniel Jr., R. (1996a), "Canonical URC Representation and Operations."
`http://www.acl.lanl.gov/URC/cspec.html.`

Daniel Jr., R. (1996b), "URC Storage Facilities and Transformation Operations."
`http://www.acl.lanl.gov/URC/.`

Daniel Jr., R. and C. Lagoze (1997), "Extending the Warwick Framework – from Metadata Containers to Active Digital Objects," *D-Lib Magazine*, November.
`http://www.dlib.org.`

Dempsey, L. (1996), "Roads to Desire," *D-Lib Magazine*, August.
`http://www.dlib.org.`

DESIRE, "The DESIRE Project."
`http://www.nic.surfnet.nl/surfnet/projects/desire/.`

Deutsch, P., A.E. Bunyip, M. Koster, and Stumpf (1995), "Publishing Information on the Internet with Anonymous FTP."
`http://info.Webcrawler.com/mak/projects/iafa/iafa.txt.`

DIF (1996), Global Change Master Directory, Directory Interchange Format (DIF) Writers Guide, version 5.
`http://gcmd.gsfc.nasa.gov/difguide/difman.html.`

DIGGER, `ftp://services.bunyip.com/pub/mailing-list.`

Faltstrom, P., L.L. Daigle, and S. Newell (1998), "Architecture of the Whois++ Service."
`http://www.ietf.org./internet-drafts/draft-ietf-asid-whoispp-02.txt.`

Ferber, R. (1997), "Hypertext and Metadata."
`http://www-cui.darmstadt.gmd.de/mind/.`

FGDC, Content Standard for Digital Geospatial Metadata, Federal Geographic Data Commitee.
`http://geochange.er.usgs.gov/pub/tools/metadata/standard/metadata.html.`

GILS, Version 2 of "Application Profile for the Government Information Locate Service."
`http://www.usgs.gov/gils/prof_v2.html.`

Goodchild, M.F. (1996), "Alexandria Digital Library."
`http://alexandria.sdc.ucsb.edu/public-documents/metadata/metadata_ws.html.`

Gravano, L., K. Chang, H. Garcia-Molina, C. Lagoze, and A. Paepcke (1997), "STARTS: Stanford Protocol Proposal for Internet Retrieval and Search."

Guha, G.V. (1997), "Meta Content Framework."
`http://mcf.research.apple.com/hs/mcf/html.`

Heery, R. (1996a), "Review of Metadata Formats," Program 30/4, preprint draft.
`http://www.ukoln.ac.uk/metadata/review.`

Heery, R. (1996b), "ROADS: Resource Organisation and Discovery in Subject-Based Services."
`http://www.ukoln.ac.uk/ariadne/issue3/roads.`

HTML, "HTML 4.0, W3C's next Version of HTML."
`http://www.w3.org/MarkUp/Cougar/.`

Iannella, R. and A. Waugh (1997), "Metadata: Enabling the Internet."
`http://www.dstc.edu.au/RDU/reports/CAUSE97.`

ISBD (1997), International Federation of Library Associations and Institutions, "ISBD(G): General International Standard Bibliographic Description: Annotated Text," London, IFLA International Office for UBC, 1.

Jain, R. and A. Hampapur (1994), "Metadata in Video Databases," *Sigmod Record 23*, 4, 27–33.

Kahn, R. and R. Wilensky (1995), "A Framework for Distributed Object Services." http://WWW.cnri.reston.va.us/home/cstr/arch/k-w.html.

Kashyap,V., K. Shah, and A. Sheth (1995), "Metadata for Bulding the Multimedia Patch Quilt," In *Multimedia Database System: Issues and Research Directions*, Springer-Verlag.

Knight, J. and M. Hamilton (1996a), "Dublin Core Qualifiers," ROADS project, Department of Computer Studies, Loughborough University. http://www.roads.lut.ac.uk/Metadata/DC-Qualifiers.html.

Knight, J. (1996b), "MIME Implementation for the Warwick Framework." http://www.roads.lut.ac.uk/MIME-WF.html.

Koch, T., A. Ardo, A. Brummer, and S. Lundberg (1996), "The Building and Maintenance of Robot Based Internet Search Services: A Review of Current Indexing and Data Collection Methods." http://www.ub.lu.se/desire/radar/reports/D3.11/.

Lagoze, C. and J.R. Davis (1995), "Dienst – An Architecture for Distributed Document Libraries," *Communications of the ACM 38*, 4, 47.

Lagoze, C., C.A. Lynch, and R. Daniel Jr. (1996), "The Warwick Framework – A Container Architecture for Aggregating Sets of Metadata." http://cs-tr.cs.cornell.edu/Dienst/UI/2.0/Describe/ncstrl.cornell/TR96-1593?abstract=.

Lagoze, C. (1997), "From Static to Dynamic Surrogates – Resource Discovery in the Digital Age," *D-Lib Magazine*, June. http://www.dlib.org.

Lasher, R. and D. Chen (1995), "A Format for Bibliographic Records," RFC1807. http://ds.internic.net/rfc/rfc1807.txt.

MARBI Discussion (1995), "Mapping the Dublin Core Metadata Elements to USMARC," Paper No. 86. gopher://marvel.loc.gov:70/00/.listarch/usmarc/dp86.doc.

MDIS, "Leading the Industry Initiative for Metadata Interchange." http://www.he.net/∼metadata.

Miller, E.J. (1996a), "An approach for Packaging Dublin Core Metadata in HTML 2.0." http://www.oclc.org:5046/∼emiller/publications/metadata/minimal.html.

Miller, E.J. (1996b), "Issues of Document Description in HTML." http://www.oclc.org:5046/∼emiller/publications/metadata/issues.html.

MSE, http://www.dstc.edu.au/RDU/Harvest/Meta/about.html.

NetEc, http://cs6400.mcc.ac.uk/NetEc.html.

Netscape, http://home.netscape.com.

Nordic (1996), "The Nordic Metadata Project." http://linnea.helsinki.fi/meta/index.html.

OCLC (Online Computer Library Center Homepage), "Building a Catalog of Internet-Accessible Materials." http://www.oclc.org/oclc/man/catproj/catq-a.htm.

OMNI, http://omni.ac.uk/.

Payette, S. and C. Lagoze (1998), " Flexible and Extensible Digital Object and Repository Architecture (FEDORA)," *European Digital Library Conference*.

PICS (1996), "PICS: Internet Access Controls Without Censorship," *Communications of the ACM 39*, 10.

PICS (1998), http://www.w3.org/PICS/refcode/.

Prabhakaran, B. (1997), *Multimedia Database Management Systems*, Kluwer Academic Norwell, MA.

PURL, "Persistent Uniform Resource Locators." http://purl.oclc.org.

RDF, "Resource Description Framework Model and Syntax." http://www.w3.org/TR/WD-rdf-syntax/.

RLG (1997), "Guidelines for Extending the Use of Dublin Core Elements to Create a Generic Application Integrating all Kinds of Information Resources." http://www.rlg.org/metawg.html.

Roszkowsky, M. and C. Lukas (1998), "A Distributed Architecture for Source Discovery Using Metadata," *D-Lib Magazine*, June. http://www.dlib.org.

SAIF, "Spatial Archive and Interchange Format: Formal Definition." http://www.wimsey.com/∼infosafe/saif/SaifHome.html.

Schatz, B. and H. Chen (1996), "Building Large-Scale Digital Libraries." http://www.computer.org/pubs/computer/dli.

Seaman, D. (1994), "Campus Publishing in Standardized Electronic Formats – HTML and TEI." http://etext.lib.virginia.edu/articles/arl/dms-arl94.html.

Seligman, L. and A. Rosenthal (1996), "A Metadata Resource to Promote Data Integration," In *Proceedings of IEEE Metadata Conference*, Silver Spring, MD.

Shklar, L., K. Shah, and C. Basu (1995), "Putting Legacy Data on the Web: a Repository Definition Language," In *WWW'95 Conference*. http://www.igd.fhg.de/www/www95/papers/78/irdl.html.

SOIF, "Summary Object Interchange Format." http://harvest.cs.colorado.edu/.

SOSIG, "Social Science Information Gateway." http://sosig.ac.uk.

Swetland, A.J.G. (1996), "Encoded Archival Description Document Type Definition (DTD)," final draft. http://scriptorium.lib.duke.edu/people/tom/guidelines/.

UKOL, "Metadata – Mapping Between Metadata Formats." http://www.ukoln.ac.uk/metadata/interoperability/

URC, "Uniform Resource Characteristics (URCs)." http://www.acl.lanl.gov/URC/.

URN (The URN Implementors) (1996), "Uniform Resource Names," Progress Report, *D-Lib Magazine*, January. http://www.dlib.org.

USMA (1996), "Marbi, the USMARC Formats: Background and Principles," American Library Association's ALCTS/LITA/RUSA, Machine-Readable Bibliographic Information Commitee. http://lcweb.loc.gov/marc/.

Vancouver Webpages (1997). http://vancouver-webpages.com.

Weibel, S., J. Grodby, and E. Miller (1995), "OCLC/NCSA Metadata Workshop Report," Dublin, EUA. http://www.oclc.org:5046/oclc/research/conferences/metadata/dublin_core_report.html.

Weibel, S. (1996), "A Proposed Convention for Embedding Metadata in HTML." http://www.oclc.org:5046/∼weibel/html-meta.html.

Weibel, S. and E. Miller (1997a), "Image Description on the Internet," *D-Lib Magazine*, January. http://www.dlib.org.

Weibel, S., R. Iannella, and W. Cathro (1997b), "The 4th Dublin Core Metadata Workshop Report," DC-4, National Library of Australia, Canberra, *D-Lib Magazine*, June. http://www.dlib.org.

Weibel, S. and J. Hakala (1998), "DC-5: The Helsinki Metadata Workshop. A Report on the Workshop and Subsequent Developments," *D-Lib Magazine*, February.

XML, "Meta Content Framework Using XML." http://www.w3.org/XML/TR/NOTE-MCF-XML-970606.

Z3950, "Z3950: The Search and Retrieval Protocol." http://lcweb.loc.gov/z3950.