Developing a Metadata Strategy

By

Grace Agnew
Associate University Librarian for Digital Library Systems
Rutgers, the State University of New Jersey

Metadata can be simply defined as "data about data." In the traditional library setting, metadata resides in a library's catalog and describes information resources collected or managed by the library. In the digital information age, however, metadata has truly come into its own. Metadata brings intelligence and coherence to digital collections, as well as order and meaning to the fragmented web. Metadata documents every facet of a digital library initiative—the selection, organization, preservation, discovery and interpretation of digital information.

MARC (MAchine Readable Cataloging) was a tremendous breakthrough in the evolution of metadata. MARC adhered to well-documented formatting principles—the Anglo-American Cataloguing Rules and International Standard Bibliographic Description (ISBD) punctuation to create coherent, interoperable metadata that could be shared and interpreted by an international community. MARC remains the most mature metadata standard available. In recent years, many new metadata schemas have evolved, generally geared to an information format, such as Encoded Archival Description (EAD), designed for archival finding aids for a community of users, such as CIDOC, or for an information medium, such as TEI (Text Encoding Initiative) for text or MPEG-7 for multimedia information. XML-based schemas to describe and document different types of digital information--music notations, mathematical equations and electronic journals among others—have also appeared. Many of these schemas incorporate descriptive metadata in the header or even the body of the XML document. Libraries have many choices and options for rich description of digital resources.

Libraries are presented with an abundance of tools and technologies for describing and managing digital objects, but what is lacking is a road map to a coherent metadata strategy. How does a library decide how to invest its scarce human resources to best serve its primary user population? Metadata was much simpler and the opportunities for costly error when the only question was how? —how to implement MARC. Now the questions are more complex—what to implement and why?

A good metadata strategy should possess the following qualities:

- *Scalable*. Involving a flexible design that supports changing user needs and new technologies.
- *Standardized*. Shareable by users and metadata registries worldwide
- *Unambiguous*. The information contained within the metadata should provide for consistent, unique interpretation by human and machine users.
- *Effective*. The metadata must persist over space and time, be readily accessible to users and integrate tightly with the information resources it describes.
- *Integrated*. The metadata must integrate with legacy metadata, such as the library's primary catalog, to present uniform and consistent description and management of the library's information resources.

Metadata may be intrinsic to the digital object, such as the header information in a TEI-encoded text, extrinsic to the digital object, such as the metadata stored in a centralized database, or a combination of the two strategies. An example of a combination strategy would be the harvesting of intrinsic metadata from the digital object into a centralized database. Metadata can be automatically generated, particularly from intelligent digital objects, such as XML-structured documents, or human-created. What is most important, however, is that metadata is based on the needs of the primary user base and represents a solid understanding of the user's information universe. An effective metadata standard

will be designed to locate the information object within the user's information universe, so that the information object makes contextual sense within the user's overall information needs.

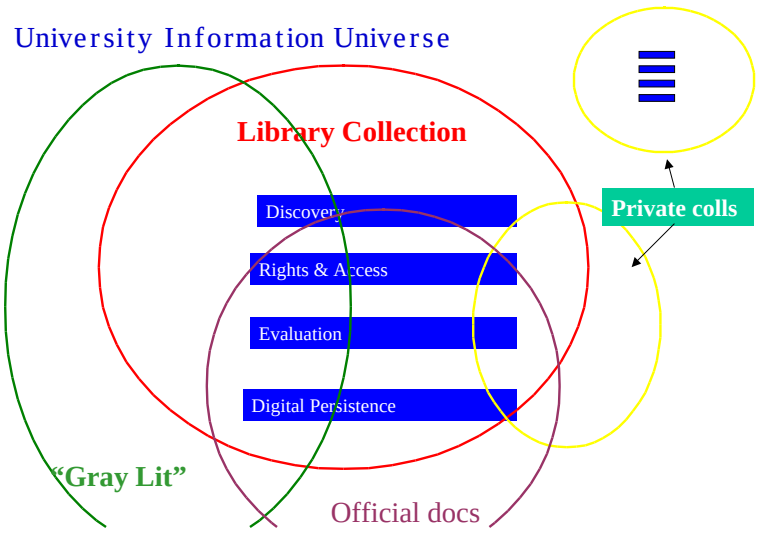
The first step—and the most difficult—is to truly understand the information universe of the primary customer base. This information universe can be understood and documented as a basic model, consisting of entities, relationships among entities and attributes of entities. The primary entities involved are the user, the metadata creator, particularly in relationship to the user, the metadata that enables discovery and evaluation of information, and the digital information objects that are being described, discovered and used within the information environment.

The information universe involves the constraints on the user that dictate what information is needed and how it is used. Common constraints include knowledge domains, such as astrophysics or medicine, organizations, such as a university or a company, and the user's application needs for information, such as research, teaching, or industrial needs. It is critical to look beyond the primary domain, to related or edge domains. A university may be part of a consortium, for example, or may be collaborating with another university in distance education courses. A study of student citations on research papers may indicate a heavy reliance on general web resources, discovered via an Internet search engine, as opposed to the library's collection. It is important to use metadata to integrate the digital collections described and made accessible by the library into the larger information domain of the customer.

A flexible metadata strategy will involve an understanding of the changing needs and roles of the user. Web-based search engines such as Google and Alta Vista have become simpler to use but increasingly sophisticated in their processing and data mining capabilities. Users have come to expect that information discovery and retrieval will be fast, contextual, customized to their unique needs, and accurate. User roles are also evolving, as web publishing becomes simpler to employ. A typical academic user is an information seeker, an information publisher and also a lifelong learner, as graduates in the commercial world struggle to keep up with rapidly changing technologies and expanding knowledge bases in every field. It is impossible to effectively grasp and utilize the exploding information core in any field—the humanities and social sciences as well as medicine and the sciences.

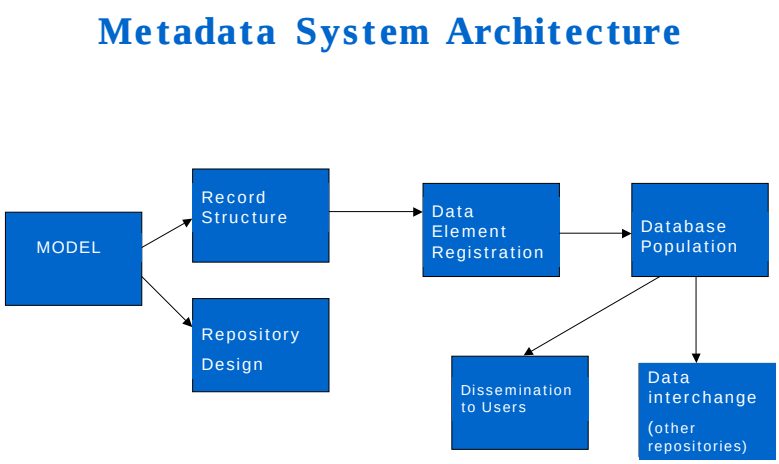
Begin your information model by understanding your user base and the information domains and information seeking behaviors of your users. Conduct interviews and surveys; study usage logs for the library's digital information sources. A good way to understand your users' information needs is to examine end products. In the university setting, this may be the textbooks and materials placed on course reserves, the books, articles, term papers, theses and dissertations produced by faculty and students, the grants applied for, and the digital information self-published by faculty and students on the web.

ata in that universe, looks



d the information resources
eeds for discovery,
y.

A metadata system will include the all the components in the architecture diagram below:



The metadata system will be based on the information model that represents your best understanding of your users, their information domain, and how they gather and use information resources. Based on the needs of your user population, you will select a primary metadata schema to either serve as the sole metadata schema or, as I recommend, as a "root" schema, to which other schema will append as extensions based on the needs of the subject, the user or the resource itself.

Selecting and implementing a "root" schema can enable metadata format independence, since a good root schema will map readily, and thus be subsumed readily, within schemas with more extensive elements or schemas that are customized for particular subjects or formats, such as the geospatial standard, CSGDM (Content Standard for Digital Geospatial Metadata, more commonly known as "FDGC"). Two root schemas worth serious consideration are MARC and Dublin Core.

MARC is currently in the 21st edition. As mentioned earlier, MARC is a mature, well-maintained standard with prescribed formatting rules and controlled vocabularies for every element. MARC is found in library catalogs around the world and is readily shared and interpreted by ILMSs (Integrated Library Management Systems) and library users worldwide. Much community effort has been expended to develop core MARC profiles, particularly the PCC (Program for Cooperative Cataloging) MARC core record. MARC maps readily to many metadata schemas, or perhaps more accurately, most metadata standards have developed mappings to this pre-eminent legacy standard.

Dublin Core, which consists of 15 optional, repeatable elements, is more properly an element set than a schema. In fact, developers and proponents of Dublin Core rarely agree on whether Dublin Core is a true "schema" or not. However, even in its unqualified "element set" form, it has been used by many digital library initiatives and consortia as a schema. The CIMI (Computer Interchange of Museum Information) is one consortia that has developed an application profile for all fifteen optional, repeatable elements, to great effect. Dublin Core is a good choice for a root schema because it was developed primarily to support mapping among schemas, because it supports data sharing initiatives such as the Open Archives Initiative (OAI) protocol and because it was designed for simplicity of use so it can enable collaboration with digital publishers and creators who are not metadata specialists.

Implementing a core or "root" schema implies that your organization will be developing an application profile for the schema, since a core schema is frequently a distillation of the most critical elements from a much more complex schema, such as MARC or, in the case of an existing core schema, such as Dublin Core, the application profile will impose order and constraints on unconstrained data elements. Developing the application profile for your root schema is one of the most important components of metadata system development.

Before developing an application profile, it is important to understand the parts of a metadata schema, as expressed within a single unit describing an information object—the metadata record. The atomic element of a data record—the smallest unit of independent semantic meaning—is the data element. The data element, often expressed as a field or subfield within a metadata record, is considered an atomic element, but the truth is that "irreducible" or "atomic" semantic meaning is determined by the metadata developers. An atomic element can be as broad or as specific as the needs of the community dictate. For example, "place of publication" can be a data element for one community, and the formatting rules can allow "city," "state," "country," or any combination of the three elements expressed in a free-text format. Alternatively, metadata for another community, such as a book distributor or book seller community, might be much more specific, with "city of publication," "state of publication," and "country of publication" each required as separate data elements with controlled vocabularies and precise formatting rules for each element.

It is not critical to develop data elements to some abstract principle of atomic distillation so much as it is critical that data elements conform to three basic principles: (1) the data elements have meaning and are useful to the primary user community; (2) the data elements are well-developed with definitions, formatting principles and controlled vocabularies, to enable interpretation and sharing by human and computer users and (3) the data elements are well-documented in a data registry.

The next component of a metadata schema are the rules that apply to each data element. These rules include the primary constraints—whether a data element is optional, mandatory, or recommended; whether a data element is repeatable; and the recommended or required order of appearance in the metadata record.

The next component of a metadata schema are the attributes of each data element, which are primarily "type," "role" (for agents, such as creators, contributors and publishers), "label," and schema--the schema to which the data element belongs or the schema used to format the data element. Different discussions of metadata may refer to these attributes by different names, but the concepts behind the attributes remain the same. Each of these attributes is discussed in turn.

"Type" is used to add qualifications, or additional definition, to data elements, such as the qualifiers "medium" and "extent" to the data element "format" in Dublin Core. Type can be used as an attribute for a data element or, depending on the specificity of your application profile or schema, "type" can create a new data element from a broader data element. So why have a "type" attribute at all? Why not just acknowledge that your data element was too broad in the first place? The reason is that your metadata architecture needs to be flexible and extensible to support a variety of purposes. Your root

metadata schema must accurately describe an information object and support initial discovery and basic evaluation of the information object by the end user. However, your root metadata schema must also flexibly map to other standards, which may be more specific than the root schema.

For example, if Dublin Core is your root schema, a data element may be "creator," which is a broadly shareable data element for data sharing initiatives such as OAI or Z39.50. It conveys adequate semantic meaning to an end user to enable the end user to discover and evaluate an information object. It maps readily to other standards, such as MARC, but not at a granular, or one-to-one mapping. A gross mapping may meet your needs initially, but as collection needs change, a finer, more granular mapping may be needed. For example, the Georgia Tech Library is beginning two digital text projects—sponsored research reports and digital theses and dissertations. Both of these collections have been cataloged for years in their analog formats in the library's MARC-based catalog. However, in their digital form, they will be cataloged in Dublin Core for searching and retrieval via digital text portals.

To insure equal accessibility between the large amount analog, shelved materials and the smaller, full-text digital collections, two strategies will be employed: federated searching across the MARC catalog and the Dublin Core database will be offered at the portal and the Dublin Core records will be exported in MARC format and loaded directly into the MARC database. Obviously, the more granular the mapping, the more meaningful the mapped record will be and the less editing will be required for the exported record. The goal is obviously to transport the record from one database to another with no loss of meaning and with minimal or no editing required for the more complex and granular MARC record. By including the "type" attribute of "PersName," "CorpName" and "ConfName" for creator and contributor, the data elements can map more readily to separate MARC elements of 100/700, 110/710 and 111/711. At the same time, this level of granularity may not be required for discovery and access within the root schema, so do not require, **for the Georgia Tech Community**, separate data elements. The emphasis in that last sentence is intended to remind you that your primary user base drives the decision-making for your metadata schema.

When determining whether to create a separate data element, such as the MARC decision to have separate data elements for persons, corporations or conferences, it is necessary to step back and take a broad view of the applications and uses the metadata system will serve. If data sharing within a larger community is important, then less granular data elements ("creator," "contributor," etc.) should be used, with the attribute "type" used to provide semantic granularity when needed for mapping or for specific collections or applications where granularity is needed. Georgia Tech's sponsored research collection always includes a "sponsor" field, for the entity providing the grant funds. "Sponsor" is almost exclusively a corporate entity, so the "type" attribute could be used to derive a browsable "corporate name" index by the search engine.

"Role" is an important but neglected attribute for "agent" data elements. Agent data elements can be broadly defined, in modeling terms, as those entities that act upon or have responsibility for the information object in some way. Examples of agents in Dublin Core include "creator," "contributor" and "publisher." Within these data elements, different roles, such as "author," "editor," "distributor," "interviewer," "producer," etc are performed. Roles are particularly critical for information objects in audiovisual formats, such as moving images or audio formats. Think, for example, of the many roles played by creators (those with primary responsibility for the intellectual content of an information object) and contributors (those with secondary or contributory responsibility) in the creation of a feature film. The creator may be the producer or the director (depending on the decision of the information community), while the contributors can include the screenwriter, the cinematographer, the music composer, the principal actors, the film editors, and, as appropriate, the author of the book from which the feature film is derived (for example, Margaret Mitchell, the author on whose book the MGM classic "Gone with the Wind" is based. For footage licensing purposes, the distributor or the rights holder

(both of which are possible roles for the data element "publisher") are as critical for licensing and royalty distribution as the original publisher.

The Georgia Tech Library uses roles to great effect to customize search, display and entry screens for different communities within the Georgia Tech information universe. For example, for the forthcoming digital theses and dissertations collection, students will be submitting their theses and dissertations with simple forms including basic Dublin Core elements of "creator," "contributor" and "publisher." However, these agent elements will be labeled by role: creator; thesis advisor; committee member and department. Search screens and displays will reflect these roles, which are more meaningful to the user community than the data elements. However, when the metadata records are shared with the Networked Library of Digital Theses and Dissertations (NLDTD) in the required Dublin Core format, the data elements without roles will be exported for maximum interoperability within a large, international consortium. As this example, and the examples for "type" demonstrate, attributes can be used with great effect to add extensibility, flexibility and personalization to your metadata schema, without sacrificing interoperability and future extensibility, as user needs and metadata schemas evolve.

"Label" is used to separate semantic meaning from the basic display label for a metadata schema. The Rutgers library is evolving a format-independent core metadata schema that will use "label" to distinguish data element use among a limitless number of metadata schema. Rutgers is evolving a metadata core that will be used to map among any metadata schema in use among the many libraries and departments engaged in digital activities at the many Rutgers University Libraries. For records expressed in the Rutgers core format, the data element name and the label will be the same. However, the records will be expressed in multiple formats by varying the schema and the label attributes. These attributes will be programmatically enabled to allow end users and metadata developers to move quickly and easily among metadata schemas to enable customized indexing, search and display while the core format serves as a unifying element among all the digital projects in the immersive Rutgers digital library environment. The Association of Moving Image Archivists are developing a union catalog—the AMIA Moving Image Gateway, with a core or root metadata standard to which all incoming records will be mapped. Schema and label attributes will be used to provide searchability, export and display in any format supported by the Gateway, as well as to support custom display and export for each participating archive. Dynamic web pages and search screens will allow each archive to have a web presence personalized to their collection. An additional attribute, record_location, will support record display in the data element order of the home database for transparent access to information, whether retrieved via the Gateway or using the archive's in-house metadata system.

The final attribute to consider is "schema." Schema is both an attribute of the data element and an attribute of the data value, discussed below. Schema is used to differentiate among labels to support multiple schemas in an extensible record and to indicate controlled vocabularies or formatting principles for data elements and data element attributes, such as "roles" which may use controlled vocabularies such as MARC-Relators or the Getty Art and Architecture Thesaurus. The 'schema" attribute has its own "attributes" or qualifiers, such as version number or online registry identifier (URN or URL).

The final component of a metadata record is the data value, the information specific to the described object that populates each data element in a metadata record for that object. For example, the creator ("data element") of this paper is Grace Agnew ("data value'). Data values should be formulated according to a controlled vocabulary (such as the Library of Congress Subject Headings) or according to a formatting principle, such as the Anglo-American Cataloguing Rules, 2nd ed. (AACR2), which dictates that the above data value is expressed in inverted last name, first name format (Agnew, Grace). Data values should include a schema attribute, which indicates how the value is formulated. Again, the schema attribute itself will have attributes for version number and the online registry for the schema. This will enable future machine processing by search engines designed to reference and utilize online registries.

Once you have determined the data elements you will use, the attributes of those data elements, the order in which the data elements will display in the primary record display format, and whether each element is repeatable, mandatory, or optional, it is time to document the application profile. Documentation should occur in an online registry format according to ISO standard 11179. ISO 11179 emerged primarily from the federal government data management environment. The primary goals of a registry formulated according to ISO 1179 are to standardize representation of the data element to enable shareability and durability (reuse) of the data element and the data values that populate the data element and to establish context and meaning for intelligent retrieval and interpretation of data. ISO 11179 consists of six parts:

11179-1 Framework for the Specification and Standardization of Data Elements
11179-2 Classification for Data Elements
11179-3 Basic Attributes of Data Elements
11179-4 Rules and Guidelines for the Formulation of Data Definitions
11179-5 Naming and Identification Principles for Data Elements
11179-6 Registration of Data Elements

ISO 11179, which is available from NISO, is currently in a state of flux as the standard is rewritten to support model-based principles. In particular, 11179-3 is being completely rewritten in a data model format. The entire standard supports not just the identification and description of data elements but the entire registration process, including the establishment of the metadata creators as an identified registration authority. A registry can be expressed in many formats. The Video Development Group (ViDe) is currently developing a registry for its application profile of Dublin Core for digital video in XML/RDF (resource description framework).

In an ISO 11179-compliant registry, each data element would receive a unique, unintelligent number to create a reusable, language-independent, machine-interpretable, data element. Other elements include the data element name, the data element label, the definition, the value domain (the contextual framework within which the values reside. For example, "U.S. state" is a location that can be used in a mail delivery context or a subject access context. Within each value domain, different formatting principles would apply (for example, the use of official state abbreviations as a controlled vocabulary for mail delivery vs. the state name spelled out as a controlled vocabulary subject heading. At a minimum, an ISO 11179-compliant registry should have the following elements:

- Unique Numeric Identifier (for reusability)
- Data element name
- Data element label
- Data element version (numeric or date or both)
- Data element definition
- Data element obligation (mandatory, optional, recommended)
- Data element repeatability
- Value Domain Name
- Value Domain Definition
- Registration Authority (Valid RA Code – developed according to ISO 11179 and registered)
- Registration Status (Recorded, Certified, Standard)
- Administration Status (In Quality Review, No Further Action, Final)

The final two elements refer to the status of the data element registration within the registry. Elements begin as proposals for evaluation and comment by the user community. Only after evaluation and revision by the community would a data element registration move from status of "recorded" and an administration status of "in quality review" to a final, standardized element available for full reusability by an international community. Obviously, a registry is going to use the identifier, definition, labels,

etc. of the creator or managing authority (such as the Dublin Core Metadata Initiative, for Dublin Core). When creating a registry, you will add information for those elements that you have added or customized for your application profile.

In addition to descriptive metadata elements, a metadata record may include administrative or meta-metadata elements. Examples include the metadata creator, date of creation, date of modification and date of deletion. This last element is very important because most metadata is stored in a database management system, many of which make no provision for storing and maintaining deleted records. Populating a "date of deletion" data element with a date value can be used to trigger the writing of a brief records to a deleted record database or table that can track record deletions away from the live database. This can be critical for supporting union catalog initiatives where records are sent to a host site automatically based on modification date. For the union catalog to be accurate, deleted records must also be identified and transmitted. Another critical administrative data element is the Archive identifier. A database can support multiple archive identifiers for independent contributors to the database, such as a consortial database with many participating institutions, such as the AMIA Moving Image Gateway.

Once your schemas, and an application profile for each schema, is established, you must decide how to create and store the metadata records. Two currently valid options include relational database management systems (or alternatively, object-relational database management systems) or online storage in a web-enabled storage and display format such as HTML or XML. HTML metadata can be stored directly within <meta> tags in the header of an HTML page. Systems currently exist that will "scrape" this metadata from a contributing page to populate a fielded relational database, thus providing both intrinsic and extrinsic metadata. A better alternative is to provide metadata in XML (extensible Markup Language). I believe that XML is currently well-known enough in the library community that a discussion of the benefits of XML over HTML is not required for this paper.

XML metadata can be provided in the header and "scraped", as described above), stored in a relational database and 'staged" for export and display in XML formats (which are dynamically repurposed as HTML displays using a style sheet, either CSS (cascading style sheets) or XSLT (XML Style Sheet Language Transformations). Newer versions of relational database management systems, such as Oracle v. 9, IBM's DB2 and the open source Zope DBMS support, or will soon support, storage of tagged XML elements. Native XML databases are also available but are currently not recommended since the tools to index them, share the data and repurpose the record display in response to queries, are not readily available or mature. Native XML databases are a "bleeding edge" technology at this time that should be studied and tested rather than actively utilized.

To effectively utilize XML as a display, storage or transport format, it is necessary to document your metadata record as an XML object through the use of a document type definition (DTD) or Schema, DTDs document how a conformant metadata record will be constructed and displayed and allow for validation against the DTD to insure conformance using primarily elements, attributes and entities. XML schemas, which are slowly replacing DTDs, use these same concepts but add modeling characteristics, object characteristics (such as attribute inheritance from a higher level element) and modularity to construct a schema that can be, in theory, combined with other schema in a modular fashion to create a "metaschema" and that provides a more object-oriented, less textual approach to XML object creation and validation. Tools for creating and validating schema are not as widespread as DTD tools. I recommend, if you are new to XML, that you begin by creating and validating a DTD and then migrate your DTD to an XML schema.

Although I noted above that the benefits of XML are largely understood, I want to close this paper with a discussion of perhaps the most important benefit—that of data transport and sharing. I have alluded several times to the Open Archives Initiative protocol (OAI). OAI is a simple, HTTP-based protocol that mines records automatically from participating metadata repositories to create union catalogs or

virtual exhibits. OAI retrieves records according to certain constraints—archive ID, datestamp (date created, date modified and date deleted) and set ID (corresponding roughly to collection or subject, as defined by the participants in the collaboration). OAI requires that Dublin Core simple (unqualified) in XML be supported as a base common denominator metadata record standard. XML, as a very flexible, customizable data storage and display standard, is ideal for transport of data within and among information communities.

The Resource Description Framework (RDF) , alluded to earlier, is an XML-based wrapper for transporting data elements that can reference metadata registries using the XML namespace for the machine-processing of data elements from distributed metadata registries. Another standard to watch is the XML-based Simple Object Access Protocol (SOAP), a data transport standard that transports the data payload in an XML format together with headers that can include programmed addressing or processing instructions that are stripped off before the payload is delivered to the end user. XML will also increasingly enable automatic metadata generation as XML-created data objects increasingly populate the web and can be mined for specific elements, which are scraped and written to metadatabases.

This is a necessarily brief look at the development of a flexible, extensible metadata system. One principle to take away from this paper is that "less is more." Developing a "core" or root metadata scheme is deceptively simple but will form the foundation of an extensible metadata architecture that enable your end users to personalize their discovery and access to the evolving, expanding universe of digital information.

Standards Referenced in this Paper:

Anglo-American Cataloguing Rules http://www.nlc-bnc.ca/jsc/

CIDOC – Museum Information Reference Model http://www.cidoc.icom.org/

CSGDM (Content Standard for Digital Geospatial Metadata, more commonly known as "FDGC").

CSS (Cascading Style Sheets) http://www.w3.org/Style/CSS/

Dublin Core http://www.dublincore.org

EAD (Encoded Archival Description) http://www.loc.gov/ead/

HTML (Hypertext Markup Language) http://www.w3.org/MarkUp/

ISBD (International Standard Bibliographic Description) http://www.ifla.org/VI/3/nd1/isbdlist.htm

MARC (MAchine Readable Cataloging) http://www.loc.gov/marc/

ISO 11179 http://xw2k.sdct.itl.nist.gov/l8/document-library/projects/11179-Revision/

MPEG-7 (Motion Picture Experts Group) http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm

Open Archives Initiative (OAI) Protocol http://www.openarchives.org/

PCC Program for Cooperative Cataloging) MARC Core Record
http://lcweb.loc.gov/catdir/pcc/pcc.html

RDF (Resource Description Framework) http://www.w3.org/RDF/

SOAP (Simple Object Access Protocol) http://www.w3.org/TR/SOAP/

TEI (Text Encoding Initiative) http://www.tei-c.org/

XML (Extensible Markup Language) http://www.w3.org/XML/

XSLT (XML Style Sheet Language Transformations) http://www.w3.org/TR/xslt