OCLC/NCSA Metadata Workshop Report

Stuart Weibel, Jean Godby, Eric Miller
Office of Research, OCLC Online Computer Library Center, Inc.
Ron Daniel
Advanced Computing Lab, Los Alamos National Laboratory

1.0 Executive Summary

The March 1995 Metadata Workshop, sponsored by the Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA), convened 52 selected researchers and professionals from librarianship, computer science, text encoding, and related areas, to advance the state of the art in the development of resource description (or metadata) records for networked electronic information objects.

1.1 Goals

Goals of the workshop included (1) fostering a common understanding of the needs, strengths, shortcomings, and solutions of the stakeholders; and (2) reaching consensus on a core set of metadata elements to describe networked resources.

1.2 Scope

The size and complexity of the resource description problem required limiting the scope of deliberations. Given that the majority of current networked information objects are recognizably "documents", and that the metadata records are immediately needed to facilitate resource discovery on the Internet, the proposed set of metadata elements (The Dublin Core) is intended to describe the essential features of electronic documents that support resource discovery. Other important metadata elements, such as those describing cost accounting or archiving information, were excluded from consideration. It was recognized that these elements might be included in a more complete record that could be derived from the Dublin Core by a well-defined extension.

1.3 The Intended Niche

The Dublin Core is not intended to supplant other resource descriptions, but rather to complement them. There are currently two types of resource descriptions for networked electronic documents: automatically generated indexes used by locator services such as Lycos and WebCrawler; and cataloging records, such as MARC, created by professional information providers. Automatically generated records often contain too little information to be useful, while manually generated records are too costly to create and maintain for the large number of electronic documents currently available on the Internet. Records created from the Dublin Core are intended to mediate these extremes, affording a simple structured record that may be enhanced or mapped to more complex records as called for, either by direct extension or by a link to a more elaborate record.

1.4 Next Steps

The work of the 1995 workshop is one of a series of steps being taken to improve the description of networked information objects. A Metadata Workshop Steering Committee is being formed to extend the work of the March 1995 Workshop through a series of similar activities that will bring together stakeholder communities to focus on discrete parts of the larger problem. As with the initial workshop, promoting effective communication among the communities will be a primary goal. The diversity of implementation efforts influenced by the Dublin Core testifies to the benfits of such communication, and promises to help integrate related activities among librarians, the Internet Engineering Task Force (IETF), text encoding researchers, and other researchers who have substantial investments in resource description.

2.0 Introduction

The explosive growth of interest in the Internet and the World Wide Web in the past five years has created a digital extension of the academic research library for certain kinds of materials. Valuable collections of texts, images and sounds from many scholarly communities--collections that may even be the subject of state-of-the-art discussions in these communities--now exist only in electronic form and may be accessible from the Internet. Knowledge regarding the whereabouts and status of this material is often passed on by word of mouth among members of a given community. For outsiders, however, much of this material is so difficult to locate that it is effectively unavailable.

Why is it so difficult to find items of interest on the Internet or the World Wide Web? A number of well-designed locator services, such as Lycos [http://lycos.cs.cmu.edu/] are now available that automatically index every resource available on the Web and maintain up-to-date databases of locations. But it has not yet been demonstrated that indexes contain sufficiently rich resource descriptions, especially if the location databases are large and span many fields of study. Moreover, a huge number of resources on the Internet have no description at all beyond a filename which may or may not carry semantic content. If these resources are to be discovered through a systematic search, they must be described by someone familiar with their intellectual content, preferably in a form appropriate for inclusion in a database of pointers to resources. But current attempts to describe electronic resources according to formal standards (e.g, the TEI header [TEI] or MARC [MARC] cataloging) can accomodate only a small subset of the most important resources.

Another solution, not yet implemented, that promises to mediate these extremes involves the manual creation of a record that is more informative than an index entry but is less complete than a formal cataloging record. If only a small amount of human effort were required to create the record, more objects could be described, especially if the author of the resource could be encouraged to create the description. And if the description followed an established standard, only the creation of the record would require human intervention; automated tools could discover the descriptions and collect them into searchable databases.

What should this hypothetical description contain? Put in a convenient jargon, the question is about **metadata**--literally, data about data--or the contents of a surrogate record that characterize an object. Thus the question can be recast more precisely: how can a simple metadata record be defined that sufficiently describes a wide range of electronic objects? Recognizing the need to answer this and a multitude of associated questions, the Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA) sponsored the invitational Metadata Workshop on March 1-3, 1995, in Dublin, Ohio. Fifty-two librarians, archivists, humanities scholars and geographers, as well as standards makers in the Internet, Z39.50 and Standard Generalized Markup Language (SGML) communities, met to identify the scope of the problem, to achieve consensus on a list of metadata elements that would yield simple descriptions of data in a wide range of subject areas, and to lay the groundwork for achieving further progress in the definition of metadata elements that describe electronic information. This paper reports on the progress made at that workshop.

3.0 The Dublin Metadata Workshop

Since the Internet will contain more information than professional abstractors, indexers and catalogers can manage using existing methods and systems, it was agreed that a reasonable alternative way to obtain usable metadata for electronic resources is to give authors and information providers a means to describe the resources themselves, without having to undergo the extensive training required to create records conforming to established standards. As one step toward realizing this goal, the major task of the Metadata Workshop was to identify and define a simple set of elements for describing networked electronic resources. To make this task manageable, it was limited in two ways. First, only those elements necessary for the discovery of the resource were considered. It was believed that resource discovery is the most pressing need that metadata can satisfy, and one that would have to be satisfied regardless of the subject matter or internal complexity of the object.

The discussion was further restricted to the metadata elements required for the discovery of what were called **document-like objects**, or **DLOs** by the workshop participants. It was believed that DLOs are still the most common type of resource sought in the Internet and that whatever solution could be proposed for DLOs could be extended to other kinds of resources. More importantly, the likelihood of making progress on this challenging problem would be increased if attention could initially be restricted to something familiar.

DLOs were not rigorously defined, but were understood by example. For example, an electronic version of a newspaper article or a dictionary is a DLO, while an unannotated collection of slides is not. Of course, the crux of the problem is that in a networked environment, DLOs can be arbitrarily complex because they can consist of text with callouts to images, audio or video clips, or to other hypertext documents. The participants of the Metadata Workshop made no attempt to limit the complexity of DLOs, except possibly to say that the intellectual content of a DLO is primarily text, and that the metadata required for describing DLOs will bear a strong resemblance to the metadata that describes traditional printed texts.

As a result of the restricted focus of the workshop, certain issues required for a complete description of DLOs, such as cost, archival status and copyright information, were eliminated from the scope of the discussion. Elements required for the description of objects other than DLOs, such as the elements required for the description of complex geological strata in a geospatial resource, were also beyond the scope of the discussion. The goal was to define a small, universally understood set of metadata elements that would allow authors and information providers to describe their work and to facilitate interoperability among resource discovery tools. But because the core elements do not yield a description of objects that meets the needs of specialized user communities, careful consideration was also given to mechanisms for extending the element set.

The primary deliverable from the workshop was a set of thirteen metadata elements, named the **Dublin Metadata Core Element Set** (or Dublin Core, for short) by the workshop participants. The Dublin Core was proposed as the minimum number of metadata elements required to facilitate the discovery of document-like objects in a networked environment such as the Internet. The syntax was deliberately left unspecified as an implementation detail. The semantics of these elements was intended to be clear enough to be understood by a wide range of users.

Below is an introductory discussion of the elements in the Dublin Core

Element Description

- **Subject:** The topic addressed by the work
- **Title:** The name of the object
- **Author:** The person(s) primarily responsible for the intellectual content of the object
- **Publisher:** The agent or agency responsible for making the object available
- **OtherAgent:** The person(s), such as editors and transcribers, who have made other significant intellectual contributions to the work
- **Date:** The date of publication
- **ObjectType:** The genre of the object, such as novel, poem, or dictionary
- **Form:** The data representation of the object, such as Postscript file or Windows executable file
- **Identifier:** String or number used to uniquely identify the object
- **Relation:** Relationship to other objects
- **Source:** Objects, either print or electronic, from which this object is derived, if applicable
- **Language:** Language of the intellectual content
- **Coverage:** The spatial locations and temporal durations characteristic of the object

To make this discussion concrete, consider an electronic record created with the relevant portions of the Dublin Core, and a sample syntax, that describes an electronic version of Maya Angelou's poem "On the Pulse of Morning". This description is based on a record created by the University of Virginia Library's Electronic Text Center. (For a description of that project, see Gaynor [Gaynor] .)

- **Title:** On the Pulse of Morning
- **Author:** Maya Angelou
- **Publisher:** University of Virgina Library Electronic Text Center
- **OtherAgent:** Transcribed by the University of Virginia Electronic Text Center
- **Date:** 1993
- **Object:** Poem
- **Form:** 1 ASCII file
- **Source:** Newspaper stories and oral performance of text at the presidential inauguration of Bill Clinton
- **Language:** English

Although the goal of the Dublin Metadata Workshop was to develop a simple set of metadata elements, these elements also had to be defined in such a way that they could be mapped into more complex and highly controlled systems such as USMARC. These conflicting demands have been reconciled in two ways. The first was to create a set of metadata elements with definitions that could be understood easily without the need for user training, as well as an approach to modification that meets the needs of specialized communities for more precise descriptions. This is described in Section 5. The second was to provide mechanisms for extending the core element set to describe items other than document-like objects. These mechanisms are discussed in Section 6.

As Guenther [Guenther]points out, a small set of metadata elements such as the Dublin Core would be valuable for at least four reasons. First, it would encourage authors and publishers to provide metadata, in a form that automated resource discovery tools could collect it, when they make data available. Second, it would encourage the creation of network publishing tools that contain a template for metadata elements, further simplifying the task of creating metadata records. Third, a record created with the Dublin Core could serve as the basis for a more detailed cataloging record if the need arises. Finally, if something like the Dublin Core became a standard, metadata records could be understood across user communities.

Since the difficult work of identifying a simple but useful specification for the description of networked resources has only begun, the major accomplishment of the Metadata Workshop was to define the problem and sketch out a solution. Many details of that solution need further refinement. Accordingly, one goal of this paper is to focus discussion by identifying the areas that need the most work. Section 7 lists the most pressing unresolved issues. The appendices identify further problems that arise when an attempt is made to formalize what has been proposed so far. Section 8 describes projects already underway that increase our understanding of how resources on the Internet should be described and show what is possible if a metadata element set such as the Dublin Core becomes a standard. Finally, Section 9 outlines the steps that are being taken to ensure that progress on this important problem continues to be made.

4.0 Underlying Assumptions

The discussions at the Metadata Workshop revealed several principles that should guide the further development of the element set. Adherence to these principles increases the likelihood that the core element set will be kept as small as possible, that the meanings of the elements will be understood by most users, and that the element set will be flexible enough for the description of resources in a wide range of subject areas. These principles are intrinsicality, extensibility, syntax-independence, optionality, repeatability and modifiability.

4.1 Intrinsicality

The Dublin Core concentrates on describing intrinsic properties of the object. Intrinsic data refer to the properties of the work that could be discovered by having the work in hand, such as its intellectual content and physical form. This is distinguished from extrinsic data, which describe the context in which the work is used. For example, the "Subject" element is intrinsic data, while transaction information such as cost and access considerations are extrinsic data. Though extrinsic data may be important for a complete description of an object, it is handled by the extension mechanisms described in Section 6.

4.2 Extensibility

In addition to its use in dealing with extrinsic data, the extension mechanism will allow the inclusion of intrinsic data for objects that cannot be adequately described by a small set of elements.

Extensibility is important because users may wish to add extra descriptive material for site-specific purposes or specialized fields. In addition, the specification of the Dublin Core itself may change over time, and the extension mechanism allows revisions while maintaining some backward compatibility with the originally defined element set.

4.3 Syntax-Independence

Syntactic bindings are avoided because it is too early to propose formal definitions and because the Dublin Core is intended to be eventually used in a range of disciplines and application programs. The examples in Appendix I show some sample encodings.

4.4 Optionality

For two reasons, all the elements are optional. First, the Dublin Core may eventually be applied to objects for which some elements have no meaning. Who is the author of a satellite image? Second, it seems futile to mandate complex descriptions if the creators of the content are expected to provide the descriptive material. A simple description is better than no description at all.

4.5 Repeatability

All elements in the Dublin Core are repeatable. For example, multiple author elements would be used when a resource has multiple authors.

4.6 Modifiability

Each element in the Dublin Core has a definition that is intended to be self-explanatory. However, it is also necessary that the definitions of the elements satisfy the needs of different communities. This goal is accomplished by allowing each element to be modified by an optional qualifier. If no qualifier is present, the element has its common-sense meaning; otherwise, the definition of the element is modified by the value of the qualifier.

Qualifiers will be typically derived from well-known conventions in the library community or from the field of knowledge appropriate to the resource. Qualifiers are important because they give the Dublin Core a mechanism for bridging the gap between casual and sophisticated users. For example, the data in the **Subject** element consists of any word or phrase that describes the object's content. However, a professional cataloger may wish to supply the name of the authoritative source from which the subject terms are taken. In such a case, the element may be written as **Subject (scheme=LCSH)**, indicating that the subject terms are taken from the Library of Congress Subject Headings.

5.0 A Detailed Description of the Elements in the Dublin Core

This section presents detailed definitions of the elements in the Dublin Core. The simple and informal records that would result from the application of this element set will help raise the standards of resource description without imposing the overhead normally associated with more exacting standards. But when necessary, the elements in the Dublin Core should support definitions that are precise enough to enable the mapping of records to widely used standards such as USMARC, TEI or FGDC.

To accomplish this goal, each element in the Dublin Core can be qualified with a scheme. Schemes are used whenever it is necessary to describe the rationale for the encoding of the data associated with the element, such as when reference is made to controlled vocabulary, a well-known notation or a published standard. This section describes the use of schemes, but since qualification is still an active topic of discussion, other qualifiers will undoubtedly be proposed in future versions of the Dublin Core. See

Appendix II for a sample SGML Document Type Definition (DTD) [Herwijnen] that defines additional qualifiers for some of the elements.

A comment about syntax is necessary at this point. As stated in Section 4.3, the elements in the Dublin Core were purposely defined in a syntax-independent manner. Nonetheless, providing examples is often the best way to communicate, so the following descriptions include examples represented in a simple syntax that might be suitable for a HyperText Markup Language (HTML) document. This should not be considered as the definitive representation, but rather as a possible representation.

5.1 Subject

The **Subject** is the field of knowledge to which the work belongs. This may be a general description of a broad discipline, or a series of descriptors of differing scope.

The **Subject** element can be qualified by a scheme, which specifies adherence to a known classification system such as the Library of Congress Subject Headings, the Dewey Decimal System, or the Art and Architecture Thesaurus, to name a few. For example:

- Subject (scheme=LCSH) = UNIX (Computer system)
- Subject (scheme=Dewey Decimal System)=004.251 Supercomputers--systems design

Without the scheme, the **Subject** element is a keyword and may contain any word or phrase that describes the intellectual content of the object. For example:

- Subject = Network Information Discovery and Retrieval
- Subject = Information Retrieval
- Subject = Cataloging of Internet Resources
- Subject = Metadata
- Subject = Description of electronic resources

5.2 Title

Title is defined as the name of the object. Most document-like objects will have an obvious title, but if the Dublin Core is eventually used to describe resources such as satellite images or objects in a museum, there will be no designated character string that is understood as a title. A descriptive phrase might be appropriate instead. For example:

- Title = Statistical Summary of Web Site Usage for NASA’s Shoemaker Levy Home Page
- Title = A Description of Networked Information: The Dublin Core Element Set
- Title = Benjamin Franklin’s Spectacles

If greater precision is desired, titles may be qualified by a scheme. For example:

- Title (scheme=AACR2) = The structure of language

5.3 Author

The common understanding of **author** is the person(s) or agent primarily responsible for the intellectual content of the work. For example:

- Author = Smith, Jeffrey K.
- Author = World Book Encyclopedia

The **author** element could also be modified by a scheme such as USMARC. In the example below, the author is a person with an honorific title who lived between 1859 and 1930.

- Author (scheme=USMARC) = 100 1 Doyle, Author Conan $c Sir, $d 1859-1930

This example illustrates an important problem. As Guenther [Guenther] argues, the Dublin Core does not achieve an unambiguous mapping to a complex scheme such as USMARC with the above syntax because the **Author** element can map to many fields in the USMARC standard. It is also necessary to specify the portion of the external scheme that is being referenced. This is a general problem with all of the Dublin Core elements, and more discussion is required to solve it in a way that is satisfactory to all affected user communities.

5.4 Publisher

Publisher is defined as the the agent or agency responsible for making the object available. For example:

- Publisher = University of Virginia Center for Electronic Text

5.5 OtherAgent

OtherAgents are the persons or organizations other than authors or publishers who have made significant intellectual contributions to the work. Strictly speaking, authors, publishers, and others responsible for intellectual content could all be described by an appropriately qualified element such as **OtherAgent**. The more verbose description adopted here is intended to make clearer the common roles of Author and Publisher. The **OtherAgent** element is intended to describe less common roles, such as editor, illustrator, compiler, convenor, photographer, or any secondary but significant role of content responsibility.

OtherAgent can be further specified using a scheme. For example, the TEI [TEI] standard contains valuable detail regarding the roles involved in creating document-like objects.

- OtherAgent (scheme=TEI) = <pubStmt><role>Creation of machine-readable version</role> <name>The Ohio State University</name> </pubStmt>

Some workshop participants suggested that it would be useful to define a set of common roles for the **OtherAgent** element so they could be used without referring to an external scheme. If this list were defined, the OtherAgent element could take the following forms:

- OtherAgent (role=Editor) = Weibel, Stuart L.
- OtherAgent (role=Illustrator) = Dennis, Wesley

5.6 Date

The **Date** of publication is intended to reflect the date at which the object became available in its current form.

- Date = May 6, 1995

The scheme may be used to identify the syntactic form of the date. For example:

- Date (scheme=ANSI X3.30-1985) = 950506

One problem with the **Date** element is that it is potentially misleading. Since electronic objects can be easily copied, the date stamp of a particular object may have no significance. It may be necessary to define a way to refer to more meaningful dates, such as latest date of the current controlled version for executable software, or the date of the original for digitized texts. More discussion is required to resolve this issue.

5.7 ObjectType

ObjectType is defined as the abstract category or genre of the object, such as novel, poem, dictionary, thesaurus, executable software, source code, data file, or any other category judged to be useful for the discovery and retrieval of the resource being described. For example:

- ObjectType = Internet Draft Informational RFC

The scheme can be used to indicate that the words used to describe the **ObjectType** are taken from controlled vocabulary. For example:

- ObjectType (scheme=AACR2) = computer file

5.8 Form

Form is defined as the data representation of the object and is intended to provide an information seeker with information about the hardware or software resources necessary to display or operate the object. Information included should provide a sufficient description to make possible a judgement of usefulness prior to accessing an object. Examples might include Postscript-II document, Windows 3.1 executable file, HTML file, or WordPerfect 6.1 document. In the example below, the form is an Internet Media Type.

- Form (scheme=IMT) = text/html

The use of the **scheme** qualifier for the **Form** element is analogous to that defined for the **ObjectType** element.

5.9 Identifier

The **Identifier** is the string or number used to uniquely identify an object. To enhance the usefulness of this data, a scheme value will specify the authority of the identifier. For example:

- Identifier (scheme=ISBN) = 0-8230-2355-9
- Identifier (scheme=URL) = http://www.oclc.org/metadata.html

Non-public identifiers, such as a university department's technical paper number, could also be used.

5.10 Relation

The **Relation** element identifies the object's relationship to other objects, print or electronic, such as other parts of a document hierarchy, other parts of a collection of documents, or another of a series of documents.

The **Relation** element is designed to give the author flexibility when selecting the scope of the resource being described. In a hypertext environment, the resource might reasonably be a paragraph from a larger text, an entry on a Web page, or a slide in a collection. Of course, the resource may also be isolated or freestanding, in which case the **Relation** element is not relevant to the description. The simple record of the Maya Angelou poem in Section 2 omits the **Relation** element for this reason.

The exact form of the **Relation** element needs to be worked out by further discussion, but it needs to contain at least two sub-elements: a description of the relation and a pointer to the related item. In the example below, it is assumed that the sub-elements **Type** and **Identifier** have been defined. The example below contains a URL, establishing the fact that the object being described is part of the proceedings from the Third International World-Wide Web Conference.

- Relation (type=ContainedIn) (identifier = URL) = http://www.elsevier.nl/

5.11 Source

The **Source** element identifies the object from which the object being described is derived. This element is intended to connect an electronic object with a previous version, possibly in another medium, that establishes the history of the object. The **Source** element identifies other objects with the same intellectual content as the resource being described, while the **Relation** element identifies objects of different intellectual content with which the resource is logically connected.

The data in the **Source** element is most easily understood as an identifier, named in the scheme, that uniquely points to the record describing the previous version. For example:

- Source (scheme=ISBN) = 0-201-63337-X

If it is more convenient, the **Source** element may contain an extensive bibliographic citation or even a recursive instance of the Dublin Core, as one of the examples in Appendix I illustrates. In such a case, the scheme is "Dublin Core."

5.12 Language

The **Language** element should specify the language of the intellectual content of the object being described. For example:

- Language = English

If an abbreviation is used, the scheme can be used to identify the source. For example:

- Language (scheme=USMARC) = Eng

5.13 Coverage

The **Coverage** element describes the spatial and temporal characteristics of the object and is the key element for supporting spatial or temporal range searching on document-like objects that describe geospatial data The first example below is a simple, nontechnical use of the **Coverage** element. The second example illustrates spatial coverage, with the scheme identifying the data as latitude/longitude coordinates; and the third example shows temporal coverage, with the scheme identifying the syntax of the dates.

- Coverage (type = spatial) = The Atlantic Ocean
- Coverage (type = spatial, scheme = LATLONG) = {West = 180, East = 180, North = 90, South = 90}
- Coverage (type = temporal, scheme = ANSI X3.30-1985) = {Begin = 19910101, End = 19930601}

For greater precision, some workshop participants suggested that **Coverage** should be modified by the qualifiers **spatial** and **temporal**. See the sample SGML Document Type Definition in Appendix II.

6.0 Extensions to the Dublin Core

Extensibility is an essential feature of any metadata system because a single set of metadata elements, no matter how large, cannot possibly accommodate all resource types. But increasing the size of the core element set would complicate rather than simplify the problem because a large element set would be less comprehensible to a diverse user population. To reconcile these conflicting demands, the Dublin Core has been designed from the outset to be small but extensible.

This is manifested in three ways. First, local additions are accommodated by allowing elements to be added to the record that describes a resource. These additional fields are not guaranteed to be understood outside the community that proposed them, but they need not cause errors for systems that understand the core element set. The extensions may be an unstructured string of text, a pointer to another record that conforms to an established standard, or even the record itself.

The second mechanism for implementing extensibility is the "scheme" sub-element, described in the previous section. The set of values for "scheme" is open-ended and unspecified in the Dublin Core because these are expected to be supplied by user communities. For some elements, such as **Subject**, the set of values will be small because only a limited number of classification hierarchies are available from which subject terms can be derived. For other elements, the set of scheme values may be large, reflecting the complexity of the resources being described. For example, the **Identifier** element may have schemes of **FTP**, **URL**, **URN**, **ISBN**, or any number of locally assigned schemes. These scheme values must necessarily reflect our current understanding of the difficult problem of name resolution for electronic objects accessible from the Internet.

The third mechanism for extensibility is the labeling of the Dublin Core itself. This can be understood as a version number and it may be changed if new elements are added to the base set or if the semantics of existing elements changes. The current version of the Dublin Core is 0.1.

This discussion describes the principles by which records created with the Dublin Core can be extended, but a separate paper is required for a more rigorous discussion that proposes implementation solutions.

7.0 Unresolved Issues

This document describes the work of a relatively small number of people who have begun to address the difficult issues arising from the need to provide descriptions of resources in order to promote their discovery in a networked environment such as the Internet. Of course, this is a large problem, and the work has only begun. One of the goals of this document is to record what has been proposed so far--in the Dublin Metadata Workshop, and in followup discussions with the participants. If this effort has succeeded, others need not spend time covering the same ground, and it will be possible to move quickly to the issues that remain. Of course, more discussion and refinement of the core element set is expected. But there are larger problems to be solved, especially those involving versions, extensibility and character sets.

7.1 Versions

The treatment of different versions of an electronic resource has not been addressed in this document because there is no consensus regarding the definition or even the vocabulary used to describe versions of electronic objects. They seem to be fundamentally different from the versions of printed materials, such as editions and printings, which have been well-described by scholars in the library community. Unlike printed versions, new electronic versions often supplant or obliterate older versions. Electronic versions also proliferate more easily, and as a result, differences between versions may be more slight.

It is therefore nontrivial to ask: when are two different versions of an electronic resource the same work? The easy answer is that they are the same if bitwise comparisons reveal them to be identical. But this criterion is too restrictive because two electronic objects, such as a document that exists in WordPerfect and ASCII formats, could have identical intellectual content but fail the test of bitwise identity.

Without a better understanding of electronic versions, it may not be possible to use the Dublin Core to create unambiguous records that describe different versions of a resource. If the creators of records judged two versions to be the same work, they would use the **Source** element to describe the earlier version; otherwise, they would record this data in the **Relation** field. This shows that the important and reasonably well-understood concept of *Edition* can be captured in either of two elements of the Dublin Core.

7.2 Extensibility

Although the principles for extending the Dublin Core are straightforward, much work remains to be done before an important question is answered: how can records be extended in such a way that meets the needs of different communities while maintaining some level of interoperability? The extended records could conform to the specifications of the Dublin Core, yet still contain an infinite variety of data types and standards. There are several ways to solve this problem but none of them are currently feasible.

First, the problem could be eliminated by requiring a single consistent transfer syntax such as SGML. But this would violate one of the underlying assumptions of the Dublin Core and would be impossible to enforce. Second, a requirement could be added to the Dublin Core stipulating that an extended record must contain a pointer to the software that understands it. This violates the rule that all elements in the Dublin Core are optional and assumes the existence of software that understands the heterogeneous records yet to be created. Third, user communities could devise and maintain their own standards. However, there is no guarantee that the independent evolution of standards for the specification of metadata to promote resource discovery in an electronic environment will produce even minimal interoperability.

7.3 Character Sets

ASCII is the most widely deployed representation of text, and in the interest of interoperability, information exchange on the Internet relies on it almost exclusively. However, the Internet reaches communities all over the world. If it is to become a significant cultural force, the needs of languages using non-ASCII character sets will eventually have to be addressed. These issues have been avoided in this document because the intent is to adopt the solutions put forth by other standards makers.

The Multipurpose Internet Mail Extensions (MIME) introduces Internet Media Types, including text representations other than ASCII. HTML, used by most World-Wide Web browsers, is a proposed Internet Media Type as well as an SGML application. In the MIME and SGML specifications, however, character representation is notoriously complex, and the two specifications are inconsistent and incompatible. The Internet Engineering Task Force (IETF), and the MIME_SGML, HTML and HyperTextTransfer Protocol (HTTP) working groups are attempting to rectify these inconsistencies and are discussing the best ways of incorporating text representations other than ASCII.

7.4 "Third-party" Metadata

In the largest sense, "metadata" includes any information which purports to be about other information. Some of the most useful metadata is produced by persons other than librarians and document owners, and it can be found neither in card catalogs nor in self-descriptions of the documents themselves. Many kinds of such "third-party" metadata (e.g., bibliographies) are indispensable aids to information discovery. It should be possible to allow topic-oriented metadata documents with semantic network functionality to be cooperatively authored, interchanged, and integrated into master documents. Such documents (and amalgamated master documents) might resemble traditional catalogs, indexes, thesauri, encyclopediae, bibliographies, etc., with functional enhancements such as the hiding of references that are outside the scope of the researcher’s interest, etc. Early work in this area is being done by the [CApH] group.

8.0 Implementations

The OCLC/NCSA Metadata Workshop is one of a series of developments that will lead to more effective resource discovery systems for the Internet. Other developments include the Denver 1992 data element discussions, the Library of Congress 1994 Workshop on the description of electronic resources, the IETF Uniform Resource Locator working group meetings, the Digital Library projects funded by the National Science Foundation, and many other activities in stakeholder communitites, as well as the experimental Web crawler "infobots."

It is no small challenge to integrate these disparate activities, each with vocabularies, agendas, and objectives of their own, into a coherent whole. One of the goals of the OCLC/NCSA Metadata Workshop was to bring many of the relevant stakeholders together for such discussions, and some progress has been achieved toward this end.

All of these activities are but the first hesitant steps toward the goal of rational resource description. Whatever advances in understanding and communication among the participants were achieved at the Dublin Workshop, the most important measures of success will be the dissemination of these ideas in the community. But the ultimate proof is in the implementations of these ideas.

A number of Metadata Workshop conferees represent organizations that have ongoing activities or are starting activities that will be influenced by the results of the workshop. These include:

8.1.1 The OCLC Spectrum Project

The primary goal of the Spectrum project is to develop a tool that enables individuals, with or without specialized knowledge of library cataloging or markup to create records for describing and accessing networked electronic resources of various types.

Contact:"Diane Vizine-Goetz" email = vizine@oclc.org

8.1.2 The OCLC Internet Resources Cataloging Project

A group of volunteer libraries is participating in a U.S. Department of Education-funded project to identify, select, and catalog Internet-accessible electronic resources using standards and conventions widely adopted in the library community.

The overall objectives of this project are to employ, evaluate and extend the library catalog model to embrace Internet resources, and to focus the intellectual resources of scores of professions on the attendant problems and opportunities. Records created in this project will conform to USMARC standards and will be suitable for integration within local and national library catalogs.

Contact: Erik Jul, *jul@oclc.org*

8.1.3 Library of Congress

The Machine-Readabale Bibliographic Information (MARBI) Committee at LC has drafted a discussion paper (DP86: Mapping the Dublin core metadata elements to USMARC [Guenther]) for review at the summer meeting. MARBI is the committee responsible for overseeing changes to the USMARC format.

Contact: Rebecca Guenther, *rgue@loc.gov*

8.1.4 O'Reilly Associates

O'Reilly & Associates, a leading publisher of Internet books, is exploring online publishing with HTML and SGML. They are supporting a definition of HTML 2.0 containing the META element to permit the inclusion of metadata records such as the Dublin Core.

Contact: Terry Allen, *terry@ora.com*

8.1.5 Los Alamos National Laboratory and Indiana University

Researchers at LANL and IU are cooperating in the implementation of a META tag implementation for application of the Dublin Core element set in HTML documents.

Contact: Ron Daniel Jr., *rdaniel@acl.lanl.gov*
Contact: Pete Percival

8.1.6 Bunyip Systems

Bunyip will be indexing the DublinCore in its deployment of WHOIS++. In addition, Bunyip will be advocating the use of the Internet Anonymous FTP Archive (IAFA) templates (a superset of the Dublin Core) for the indexing of Anonymous FTP archives through its archie service.

Contact: Chris Weider, *clw@bunyip.com*

8.1.7 Georgia Institute of Technology

Implementors are using the Dublin Core as the set of metadata elements to include in a resource discovery system based on whois++ and the centroids mesh. It will act primarily as a method for transforming different data formats. In particular it will translate from the more complex TEI header into a simpler attribute/value flat list based on the Dublin Core. This flat list will be the basis for limited information discovery.

Contact: Michael Mealling, *michael.mealling@oit.gatech.edu*, http://www.gatech.edu/iiir

8.1.8 SoftQuad

SoftQuad Panorama, the company's SGML viewer for the Web, will become metadata-aware as implementations of Dublin Core records become available. It will support, through the use of SGML Document Type Definitions, both the basic core set as well as site-specific extended versions.

Contact: Yuri Rubinsky, *yuri@sq.com*

8.1.9 Concordia University

The semantic header is a metadata structure which stores indexing information about Internet resources. This information will be stored in distributed databases and will be accessible for search using a GUI. The metadata elements contained in the semantic header have been influenced by the Metadata Workshop.

Contact: Bipin Desai, *bcdesai@cs.concordia.ca*, http://www.cs.concordia.ca/~faculty/bcdesai/cindi-system-1.0.html

9.0 Next Steps

Refinement and standardization of the metadata element set defined in this document will be an ongoing, dynamic process involving many stakeholder communities. No single forum will suffice to air all concerns and no single standard can be expected to accommodate the needs of all communities. The problem must be divided into manageable chunks and the process must engage the relevant stakeholder communities. Implicit in the present activity is the proposition that there are core elements common to many object types, and that a simple, extensible framework of such elements can be defined to support more complete resource descriptions.

The initial objective--the specification of elements for the discovery of document-like objects--can be extended in a variety of directions.

- Expansion of the Dublin Core to include other object types, such as services or collections.
- Expansion of the Dublin Core to embrace functionality other than resource discovery, such as archival control and the authentication of users and charging mechanisms.
- Establishing standardized methods for extensibility.
- Refinement of existing work. The Dublin Core is an untested approach to the description of resources that will need to be modified with experience.

OCLC and NCSA will establish a workshop series to address aspects of this agenda. A Metadata Workshop Steering Committee will be established to define topics and assure appropriate representation of stakeholders. Design groups of perhaps a dozen or fewer individuals will be solicited to prepare discussion papers to focus workshop activities. Participants will be invited based on their publicly evident accomplishments in relevant areas or by reviewed application. Workshops will be limited to 50 or fewer participants and conducted in roughly the style of the March 1995 Workshop.

Further work will be done at a Birds of a Feather (BOF) meeting planned for the Stockholm IETF Meeting in July of 1995. A BOF discussion is an initial step in the establishment of an IETF working group. The IETF serves best in the capacity of validating design work that is done by smaller, more focused groups. An IETF working group on Metadata will provide an effective way of involving the technical stakeholders in the computer and networking community that must implement and use the standards.

Finally, active promotion of results will be carried out by establishing liaison with formal associations of stakeholders. In the library community, MARC standards evolve under the guidance of the Machine-Readable Bibliographic Information Committee (MARBI), composed of representatives of the Library of Congress and other stakeholders in the library community. A close relationship should be sustained between this committee and the Metadata Work Group. Relationships should also be established with publishers, document vendors, SGML vendors and theoreticians working on the problem of text encoding. Other communities also have requirements that must be accommodated in any framework for resource description. These communities include the GIS community, Government information providers and business communication groups.

References

[CApH]
Conventions for the Application of HyTime (CApH). "Semantic Assignment Module" and "Topic Relationship Module." 1995. Graphic Communications Association, Alexandria, VA. (ftp.techno.com, pub/HyTime/CApH)

[FGDC]
Federal Geographic Data Committee. 1994. Content standards for digital geospatial metadata (June 8). Federal Geographic Data Committee. Washington, D.C.

[Gaynor]
Gaynor, Edward. 1994. "Cataloging Electronic Texts: The University of Virginia Library Experience." Library Resources and Technical Services 38(4): 403-413 (October 1994).

[Guenther]
Guenther, Rebecca. 1995. "Mapping the Dublin Core Metadata Elements to USMARC" MARBI Discussion Paper NO. 86.

[Herwijnen]
Herwijnen, Eric van, Practical SGML , Kluwer Academic Publishers, 1994

[MARC]
Network Development and MARC Standards, Office, ed. 1994. USMARC Format for Bibliographic data. 1994. Washington, DC: Cataloging Distribution Service, Library of Congress.

[TEI]
Sperberg-McQueen, C. M., and Leu Burnard, ed. 1994. Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford: Text Encoding Initiative.

Appendix 1.0: Sample Records Encoded Using the Dublin Metadata Core

This Appendix contains two sample Dublin Core records encoded in a simple syntax that might be translated easily to SGML. As in all other examples in this document, the syntax of these records is presented only for clarity of exposition.

The first example was created by a subject-matter specialist who has no library cataloging expertise. It describes an Internet Request for Comment (RFC) found on a Web page containing similar RFCs. The Relation element points to the Web page and to a similar record on the same page.

```
Subject:      IETF, URI, Uniform Resource Identifiers
Title:        A Unifying Syntax for the Expression of Names and Addresses of Objects
              on the Network as used in the World-Wide Web.
```

```
Title:       (Subtitle) Universal Resource Identifiers in WWW
Author:      Berners-Lee, T.
Publisher:   CERN
Date:        1994
Object-Type: Internet RFC
Form (scheme=IMT): text/plain
Identifier(scheme=URL): gopher://gopher.es.net:70/0R0-57601-/pub/rfcs/rfc1630.txt
Relation (type=child)(identifier=URL):  http://ds.internic.net/ds/dspg1intdoc.html
Relation (type=sibling)(identifier=URL):http://ds.internic.net/rfc/rfc1738.txt">
```

The second Dublin Core record was created by a librarian for a Postscript version of a monograph that is derived from a previous printed edition. It contains references to schemes commonly used in the library community. The link to the printed version is described in the **Source** element, here realized as a recursive instance of the Dublin Core record.

Example 1: Dublin Core record for an electronic version of an OCLC Research Report

```
Element Name                  Content

Subject:
  scheme=LCSH:                Internet (Computer network)
                              Cataloging of computer files
                              Information networks
                              Computer networks
                              Libraries--Communication systems
                              Information storage and retrieval
                              systems

Title:                        Assessing Information on the
                              Internet: Toward Providing Library
                              Services for Computer Mediated
                              Communication

Author:                       Martin Dillon
Author:                       Erik Jul
Author:                       Mark Burge
Author:                       Carol Hickey

Publisher:                    OCLC

Date:                         1994

Identifier:
   Scheme=OCLC:               155653163X

Object type:
   Scheme=AACR2:              monograph

Form:                         7 postscript files
                              1 Unix tar file

Relation:                     For a Web page listing Internet
                              accessible OCLC research
                              publications go to:
                              http://www.oclc.org/oclc/menu/resch
                              doc.htm

Language:                     English
```

```
Source(scheme=DublinCore):    Subject:
                                  scheme=LCSH:
                                     Internet (Computer network)
                                     Cataloging of computer files
                                     Information networks
                                     Computer networks
                                     Libraries--Communication systems
                                     Information storage and retrieval systems
                              Title:
                                     Assessing Information on the
                                     Internet: Toward Providing Library
                                     Services for Computer Mediated
                                     Communication

                              Author: Martin Dillon
                              Author: Erik Jul
                              Author: Mark Burge
                              Author: Carol Hickey

                              Identifier:
                                  scheme=OCLC Technical Report Number:
                                      1234567

                              Date: 1993
                              Object type
                                   Scheme=AACR2: monograph
                              Form:
                                Scheme=AACR2: 1 v. (various
                              pagings) : ill. ; 29 cm.
                              Publisher: OCLC
```

Appendix 2.0: A Proposed Document Type Definition for the Dublin Metadata Core

This sample DTD is included here to make the proposals in this paper more precise and to promote discussion among those wishing to make improvements. The attribute **Scheme** is used when reference is made to an external authority for notation or vocabulary.

As in all examples in this paper, the DTD in this appendix specifies one possible syntax for an SGML version of the Dublin core. It is for illustrative purposes only.

```
<!-- This is the ISO8879:1986 document type definition for the DublinCore URC.  -->
<!--  Note: This DTD is subject to discussion and/or modification by the
            participants of the OCLC/NSCA Metadata Workshop.
            95/20/06, eric j. miller, emiller@oclc.org                          -->


<!-- ============ Parameterizable Lists =============== -->

<!-- MeSH (Medical Subject Heading) Publication Types can be found
        <URL:http://www.sils.umich.edu/~nscherer/Medline/Table1.html>           -->

<!ENTITY        % Subject.Scheme
```

```
                "LCSH | MeSH | Sears | AAT | INSPEC | ERIC | DDC | Other" >

<!-- TEI Information can be found <URL:http://etext.virginia.edu/TEI.html>      -->

<!ENTITY        % Title.Scheme
                "AACR2 | TEI | Other" >

<!ENTITY        % Author.Scheme
                "AACR2 | TEI | Other" >

<!ENTITY        % OtherAgent.Scheme
                "AACR2 | TEI | MARC | Other" >

<!ENTITY        % Publisher.Scheme
                "AACR2 | TEI | Other" >

<!-- ANSIX3.30 ::== yyyymmdd (4 for the year, 2 for the month, 2 for the day)   -->
<!-- ANSIX3.43 ::== hhmmss.f (2 for the hour, 2 for the minute, 2 for the sec
        and 2 for the fraction of the second including the decimal point       -->
<!-- ANSIX3.51 ::==                                                             -->

<!ENTITY        % Date.Scheme
                "ANSIX3.30 | ANSIX3.43 | ANSIX3.51 | Other" >

<!ENTITY        % ObjectType.Scheme
                "NLM | Other">

<!ENTITY        % Form.Scheme
                "IMT | X.400 | Other">

<!ENTITY        % Identifier.Scheme
                "URN | URL | LCCN | ISBN | ISSN | SICI | MessageID | FPI | Other" >

<!ENTITY        % Source.Scheme
                "TEI | Other" >

<!ENTITY        % Language.Scheme
                "MARC | Other" >

<!ENTITY        % Relationship.Scheme
                "URN | URL | LCCN | ISBN | ISSN | SICI | MessageID | FPI | Other" >

<!ENTITY        % Hierarchy.Link
                "Top | Parent | Child | Sibling | Other" >

<!ENTITY        % Relationship.Type
                "Supersedes | Continues | Continued.From |
                Contained.In | Superseded.By | Cites | Extracted.From |
                Is.Part.Of | Contains | IsIndexOf | IsIndexedBy | GlossaryOf |
                Predecessor | Successor | IsDerivativeOf | Child | Parent |
                Sibling" >

<!ENTITY        % n.spacewindow
                "WestBounding, EestBounding, NorthBounding, SouthBounding" >

<!ENTITY        % n.timewindow
                "Begin | End" >


<!-- =========== Body of the DublinCore Metadata DTD == ============== -->
```

```
<!-- Element list: Subject to change -->

<!ELEMENT       DublinCore      - -
                (BaseDesc?, Extension*) >
<!ATTLIST       DublinCore      Version         CDATA                   #IMPLIED >

<!ELEMENT       BaseDesc        - -
                (Subject | Title | Author | OtherAgent | Publisher |
                Date | ObjectType | Form | Identifier | Relation |
                Source | Language | Coverage)* >


<!ELEMENT       Subject                 - -     ANY >
<!ATTLIST       Subject         Scheme          (%Subject.Scheme;)      #IMPLIED >

<!ELEMENT       Title                   - -     ANY >
<!ATTLIST       Title           Scheme          (%Title.Scheme;)        #IMPLIED >

<!ELEMENT       Author                  - -     ANY >
<!ATTLIST       Author          Scheme          (%Author.Scheme;)       #IMPLIED >

<!ELEMENT       OtherAgent              - -     ANY >
<!ATTLIST       OtherAgent      Scheme          (%OtherAgent.Scheme;)   #IMPLIED >

<!ELEMENT       Publisher               - -     ANY >
<!ATTLIST       Publisher       Scheme          (%Publisher.Scheme;)    #IMPLIED >

<!ELEMENT       Date                    - -     ANY >
<!ATTLIST       Date            Scheme          (%Date.Scheme;)         #IMPLIED >

<!ELEMENT       ObjectType              - -     ANY >
<!ATTLIST       ObjectType      Scheme          (%ObjectType.Scheme;)   #IMPLIED >

<!ELEMENT       Form                    - -     ANY >
<!ATTLIST       Form            Scheme          (%Form.Scheme;)         #IMPLIED >

<!ELEMENT       Identifier              - -     ANY >
<!ATTLIST       Identifier      Scheme          (%Identifier.Scheme;)   #IMPLIED >

<!ELEMENT       Relation                - -     ANY >
<!ATTLIST       Relation        Scheme          (%Relationship.Scheme;) #IMPLIED
                                Type            (%Relationship.Type;)   #IMPLIED >

<!ELEMENT       Source                  - -     ANY >
<!ATTLIST       Source          Scheme          (%Source.Scheme;)       #IMPLIED >

<!ELEMENT       Language                - -     ANY >
<!ATTLIST       Language        Scheme          (%Language.Scheme;)     #IMPLIED >

<!ELEMENT       Coverage                - -     ((Spatial | Temporal)+) >
<!ELEMENT       Spatial                 - -     ((WestBounding,
                                                  EastBounding,
                                                  SouthBounding,
                                                  NorthBounding)? | Place*) >
<!ELEMENT       Place                   - -     ANY >
<!ELEMENT       WestBounding            - -     ANY >
<!ELEMENT       EastBounding            - -     ANY >
<!ELEMENT       SouthBounding           - -     ANY >
<!ELEMENT       NorthBounding           - -     ANY >
<!ELEMENT       Temporal                - -     ((Begin, End)? | Time*) >
<!ELEMENT       Time                    - -     ANY >
```

```
<!ELEMENT       Begin                   - -     ANY >
<!ATTLIST       Begin       Scheme              (%Date.Scheme;)         #IMPLIED >
<!ELEMENT       End                     - -     ANY >
<!ATTLIST       End         Scheme              (%Date.Scheme;)         #IMPLIED >

<!ELEMENT       Extension               - -     CDATA >
```