Design and Implementation of the National Institute of Environmental Health Sciences Dublin Core Metadata Schema

W. Davenport Robertson
National Institute of Environmental Health Sciences
Research Triangle Park, NC, USA
robert11@niehs.nih.gov
Ellen M. Leadem
National Institute of Environmental Health Sciences
Research Triangle Park, NC, USA
leadem@niehs.nih.gov
Jed Dube
National Institute of Environmental Health Sciences/OAO Corp.
Research Triangle Park, NC, USA
dube@niehs.nih.gov
Jane Greenberg
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
janeg@ils.unc.edu

Abstract

The National Institute of Environmental Health Sciences (NIEHS) has formed a team to design and implement a Dublin Core-based metadata schema to enhance the public's ability to retrieve pertinent public health information on the organization's Web site. The team decided to use the DC schema because it is a de facto standard and because of its flexibility. With a little customization the team has created an NIEHS-DC metadata schema. Using this schema, Web page content creators can produce metadata that is then stored in XML files.

Keywords: *Dublin Core, Metadata, Schema, NIEHS, XML, Environmental Health.*

1. Introduction

The National Institute of Environmental Health Sciences (NIEHS) is one of 25 Institutes and Centers of the National Institutes of Health (NIH), which is a component of the U.S. Department of Health and Human Services (DHHS). The mission of NIEHS is to reduce the burden of human illness and dysfunction from environmental causes. We rely on the World Wide Web as a primary communication tool to inform the American public about breakthroughs in environmental health. Metadata is a key aspect of this tool that will enhance the public's ability to locate and retrieve the scientific information needed to improve public health.

NIEHS has approximately 25,000 Web pages for people ranging from NIEHS scientists and staff to the general public, other scientists and children. (See http://www.niehs.nih.gov/.) When people conduct searches on environmental health topics using the popular search engines on the Web, they often do not retrieve the most appropriate Web pages at NIEHS. The circumstances are frequently the same for persons searching the NIEHS Web site via the institution's internal engine. In this setting, searchers are often inundated with numerous highly technical pages that are intended for scientists or with pages that casually mention the search topic. This is the problem NIEHS sought to remedy through the implementation of the NIEHS-Dublin Core metadata schema. Stuart Weibel, in a summary report of the first Dublin Core workshop, reinforces this solution when he states, "Resource discovery is the most pressing need that metadata can satisfy."[1]

2. Search and retrieval problems at NIEHS

An organization like NIEHS has limited control over which Web pages the popular Web search engines retrieve and place at the top of their results list. Major Web search engines and catalogs utilize software agents (spiders, robots, crawlers, etc.) to

https://doi.org/10.23106/dcmi.952106581

index the entire Web continuously. Programmers of these agents are challenged to keep them moving quickly, spending little time on individual pages, yet gathering as much useful data as possible. Every crawled Website must in turn somehow cooperate with these indexing agents, essentially "handing over" metadata, *without really knowing* what the indexing agent is gathering or how it operates. Fortunately, some basic facts about the search indexers' approach are public (see Search Engine Watch, http://searchenginewatch.com/), including whether they read metatags and the relative weight attached to the metadata that is discovered in those metatags. By ensuring the presence of metadata in its Web pages, an organization should be able to improve the chances that its pages will be retrieved by those search engines that consider metatags.

NIEHS Web page development is a widely distributed effort, and there has been limited control over Web developers, including their use of metatags. The resulting quality of Web pages, in terms of how well the content is organized and prioritized, is quite variable. To improve retrieval, Web administrators have asked the Web developers to be certain that their Web page HTML title and heading tags are accurate and specific, but that still hasn't improved retrieval noticeably.

In working towards improved retrieval, NIEHS has implemented Ultraseek Server V.3.1 for indexing the NIEHS Web pages and for providing search and retrieval capability to anyone who accesses the Website. It took about one year to fine-tune Ultraseek so that it would successfully index and seamlessly merge multiple collections of documents in several different formats. An adjustment to the Ultraseek indexing algorithm resulted in the ability to give more weight to text within specified tags (e.g., title, keyword metatags, alt tags) than to text in the body of the Web pages. However, the fact remained that Ultraseek retrieved documents whose usefulness and relevancy ranking were questionable for many users. If at least the most important Web pages had metadata, retrieval would be improved, and that is one reason NIEHS embarked on this project.

A key step taken was to determine to what extent NIEHS Web developers had included metadata in their pages. Using Metabot, an inexpensive metatag scanning, discovery and insertion tool from Watchfire Corporation, an inventory of existing Web pages (excluding the NIEHS Environmental Health Information Service pages and the online journal *Environmental Health Perspectives* which have metatags) was conducted. The inventory revealed that only ten to fifteen percent of NIEHS pages contained viable keyword metatags, and that some of those were generic, copied from a single, all-purpose metatag. The inventory also identified a variety of other errors and shortcomings in existing metatags, such as key terms or words missing, technical (or lay) versions of terms missing, misspellings, unnecessary or inappropriate pluralizations or capitalizations, etc. To alleviate these problems, a collaborative team composed of librarians and Web specialists was formed at NIEHS with the intent of implementing a metadata project.

3. Collaboration and project design

Networked communication has facilitated collaboration on many different levels, such as between librarians and systems developers, between government agencies and contractors, and between research institutions and academia. It has enabled different organizations with similar interests to collaborate in pursuing a project. To address the limitations of Web search engines, the librarians at NIEHS made the decision in 1999 to embark on a collaborative project to improve search precision through the use of metadata. For technical assistance they asked the Web Services group leader from the Information Technology Support Services Contract at NIEHS to join the project team.

NIEHS has had an established relationship with the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill in the form of an internship program. This program has been going on for more than 15 years; under it, UNC has been providing library support services to NIEHS in the form of three student interns every year. Also, UNC provides guidance to the students and consultant services to the NIEHS Library staff. A logical outgrowth of this relationship was for the NIEHS librarians to request that UNC representatives join the NIEHS metadata team. For NIEHS, the benefit came in the areas of guidance on standards, a knowledge of tools to use, and manpower to help develop additional tools. For UNC-SILS, the arrangement meant an opportunity for research as well as a hands-on experience for students.

While there are a number of well-documented organization metadata projects, two informed the NIEHS project in particular. These include the Gateway to Educational Materials (GEM) project for the U. S. Department of Education[2] and the U. S. Environmental Protection Agency (EPA) metadata project (see http://www.epa.gov). The conceptual framework developed for the GEM project and also the "to do" list created for the UNC Digital Library Project (see http://www.unc.edu/projects/diglib/tasks.htm) were helpful in planning. The EPA undertaking, begun in 1996, was massive and involved the creation of a large database of metadata records, and NIEHS couldn't follow that path. Instead of creating a database of metadata records, the NIEHS metadata team decided to add the metadata to the existing Web pages. Initially, the team settled on Metabot to accomplish this, but they determined that due to the

https://doi.org/10.23106/dcmi.952106581

perceived ownership of Web pages by their content creators, a more flexible process for storing the metadata was called for, and that process involves a combination of the Dublin Core and XML.

4. The choice of Dublin Core

To improve the search engines' retrieval effectiveness, the NIEHS metadata team decided to utilize a set of descriptive metadata elements that most closely represent an existing or emerging standard. A number of communities with a vested interested in resource discovery have developed specific metadata schemas for determining not only *how* to describe electronic resources, but also *what* terminology or actual encoding of data can help to standardize the descriptive process. These schema use descriptive conventions unique to specific bodies of knowledge. However, such descriptive conventions could, while providing the means to create a corpus of rich subject-specific metadata, also serve to limit the possibility of its widespread use and adoption by other communities. In many instances, such highly specific schemes are unsuitable for the description of a broader range of resources.

NIEHS conducted an extensive investigation in order to find a metadata schema that would support descriptive access and record adequate subject and authorship metadata for resources. The subject or keyword metatag emphasis emerged because NIEHS Web logs indicate that searchers most often conduct searches using these two metadata elements. Another goal of the schemas being investigated was that they include metatags for resources which were primarily text in nature but which might also include images.

The schema chosen for NIEHS would have to succeed on many levels, but at minimum it must be a "form suitable for interpretation both by search engines and by human beings, and it must also be simple to create so that any Web page author may easily describe the contents of their page and make it immediately more accessible and more useful. As such, compromises must be made in order to provide as much useful information as possible to the searcher while leaving the technique simple enough to be used by the maximum number of people with a minimum degree of inconvenience." [3]

The first Dublin Core workshop in 1995 resulted in the development of the schema known as the Dublin Core Metadata Set. Originally consisting of a minimum "core" of 13 elements which could be used to describe a Web resource, it also provided for the addition of other elements in the future, thereby establishing its ability to grow in complexity as needs arose. The development of the Dublin Core schema was guided by the incorporation of certain principles which continue to govern its growth: "intrinsicality, extensibility, syntax independence, optionality, repeatability and modifiability". [4] Today, the Dublin Core Metadata Element Set (http://dublincore.org) consists of fifteen elements designed to provide a means for describing digital objects. It serves as a basis for categorizing and cataloging electronic resources available on the World Wide Web. As of July 2001, the DCMES has become a NISO standard and will be advanced to the American National Standards Institute for review and possible approval as an ANSI standard. (http://www.niso.org/Z3985.html).

4.1 Attributes of the Dublin Core

The NIEHS metadata team determined that the Dublin Core Metadata Element Set most closely represents a "standard" at this time, and it fulfills the following goals of the NIEHS metadata project:

1. **Stability:** Providing metadata for NIEHS Web pages requires a substantial commitment of time and effort on the part of institutional staff. The NIEHS metadata team agreed with the underlying premise of the Dublin Core Metadata Element Set, that it would always consist, at minimum, of a core set of elements, adequate for the description of document-like objects (DLOs). This philosophy insures that the NIEHS project can proceed with a set of elements that will have long-term applicability and will guarantee interoperability with other schemas.

2. **Simplicity**: One goal of the NIEHS metadata project was that metadata creation would be a combined effort of both content creators and information professionals. The Dublin Core Metadata Element Set was designed for the non-specialist to understand without any professional training. It is this underlying simplicity that was very attractive to NIEHS.

3. **Flexibility:** NIEHS requires a schema that adequately describes a range of resources and yet supports local obligatory conventions. The elements in the Dublin Core set are all optional and all repeatable. This enabled the NIEHS metadata team to determine which elements were to be mandatory, in order to assure a minimal level of description. The set of optional elements provides the means for gathering additional descriptive information, where applicable. Repeatability of certain elements is necessary in order to describe adequately, for example, multiple titles, authors, and subjects associated with NIEHS Web pages.

4. **Extensibility:** The NIEHS metadata team determined that a minimum set of mandatory elements was needed. The Dublin Core Metadata Element Set is designed to

https://doi.org/10.23106/dcmi.952106581

accommodate expansion beyond the basic 15 elements, as the needs of subject-specific groups emerge. The NIEHS metadata team created an application profile that comprises the 15 metadata elements from the Dublin Core namespace and the element of "Audience" that is part of the Dublin Core Education namespace. The long-term goal of incorporating subject terms from an existing thesaurus or controlled vocabulary will further enrich the metadata record beyond the subject terms or keywords provided by the content creators alone.

5. **Interoperability:** The selection of a schema in which the basic elements are simple, well defined and consistently applied will increase the probability of sharing data between other applications and organizations. For instance, the European Environment Agency's European Environment Information and Observation Network (see http://www.eionet.eu.int/) is using DC metadata in its Global Environmental Locator System Element Set (GELOS) (see http://www2.mu.niedersachsen.de/cds/etc-cds_neu/gelos_dc.html), and NIEHS has the potential not only to share resources with this organization, but to support cross system searching if it is desired. The NIEHS-DC application profile can serve as a model schema for other organizations with a need to create metadata for Web resources in the area of environmental health.

Other factors supporting the decision at NIEHS to use the Dublin Core element set were the wide applicability and scope of the Dublin Core Metadata Initiative itself and its stated mission to develop and promote working relationships with other standards bodies and stakeholders from other world-wide communities of interest. For purposes of the NIEHS metadata project, the team decided to incorporate all 15 of the Dublin Core Metadata Elements. And, as stated above, the team also decided to adopt the Audience element from the DC Education namespace schema.

The final NIEHS-DC metadata schema is, therefore, an application profile, comprised of the Dublin Core and the DC Education namespaces. This schema uses qualified Dublin Core, incorporating some of the qualifiers approved in July 2000 by the Dublin Core Usage Committee. (http://dublincore.org/documents/dcmes-qualifiers/). The NIEHS-DC metadata schema also incorporates the use of definitions, examples and term lists that serve as aids to the metadata creator and provide for some measure of data consistency for certain elements. The inherent flexibility built into the Dublin Core Element Set allows for the creation of metadata guided by a set of locally defined rules established by NIEHS.

4.2 The NIEHS-DC metadata schema

A number of metadata initiatives have worked with the Dublin Core schema as a core and have expanded the element set or adopted elements from other namespaces. A good example in the medical field is the French Catalogue et Index des Sites Medicaux Francophones which uses the Dublin Core as its base but omits four elements and adds eight more.[5] The GEM element set referred to earlier extends the Dublin Core Element Set by more than half a dozen elements. Both add an audience element.

Working within the Dublin Core Element Set Version 1.1, the NIEHS metadata team established the rules for using the 15 elements and the description of the kinds of data in each element. The rules had to address the needs and limitations of the content creators as well as provide a rich context in which professional catalogers could further enhance the basic metadata.

The team established a minimum standard for metadata creation by designating elements in the NIEHS-DC metadata schema as mandatory or optional. (See Figure 1.)

Mandatory Elements	Optional Elements
Title	Alternative Title
Audience	Controlled Vocabulary
Author/Contributor	NIEHS Number
Subject	Other Identifier
Publisher	Type
Date Created	Source
Date Modified	Description
URL	Relation
Language	Coverage
Format	
Rights	

Figure 1. Mandatory and optional elements in the NIEHS-DC schema.

In addition, the team determined that the following elements would be non-repeatable:

- Date Created
- Date Modified
- Description
- Rights

The NIEHS metadata team used several methods to provide clarity and guidance for the metadata creator. These methods included:

1. Re-labeling certain elements
2. Utilizing element qualifiers
3. Providing definitions and examples

https://doi.org/10.23106/dcmi.952106581

4. Creating term lists
5. Setting default values where applicable

The names of certain DC Elements were replaced with local "labels" to improve users' understanding of the use of the element within the NIEHS application. For example, the DC Element "Creator" was re-labeled as "Author/Contributor". The term "author" is readily understood in the NIEHS environment, and it is used to describe the individual responsible for the intellectual content. The inclusion of the word "contributor" in the local label expanded the "scope" of the element to enable metadata creators to name the individuals or other bodies responsible for all aspects of the resource's content. This decision was also based on the premise that Author is a subclass of contributor, a point discussed at previous Dublin Core meetings and on the DCMI listserv.

In addition, the NIEHS-DC metadata schema includes the following element qualifiers taken from the DCMES approved listing and from local qualifiers developed specifically for use in the NIEHS application:

- Alternative Title
- Controlled Vocabulary
- Date Created
- Date Modified
- NIEHS Number
- Other Identifier

They are logical refinements of the DC elements for Title, Subject, Date Created, and Identifier. The availability of these element qualifiers will help to encourage recording very specific and relevant information about NIEHS Web pages. For example, the presence of an NIEHS grant or project number is very useful to the NIEHS site, and content creators could possibly overlook them if the element qualifier were not labeled and present in the schema. The choice of descriptive labels for each of these element qualifiers further serves to clarify their intended purpose.

The NIEHS metadata team recognized that certain DC element names were adequate to describe their intended purpose, but that others could seem redundant and cause confusion. The elements which were the most difficult to define clearly for both the professional catalogers and content creators were Source, Type, Relation, and Format. To address this difficulty, the team created a set of specific definitions and examples for these elements to illustrate the type of data appropriate for each. The team also included corresponding term lists for some elements. The definitions, instructions, and term lists are as follows:

- **NIEHS-DC Metadata Element: Source**: The original source of the content of the Web page (e.g., if the information on the Web page comes from a printed source or another Web page). You should provide the identifier (URL, ISBN, ISSN, etc.) and the name of the resource from which it is derived.

- **NIEHS-DC Metadata Element: Type:** The category or genre of the resource (e.g., collection, dataset, event, image, interactive resource, etc.). Corresponding term list: Text, Image, Event, Sound, Collection, Dataset, Interactive Resource, Service, Software, Model, Physical Object.

- **NIEHS-DC Metadata Element: Relation:** An identifier of a second resource and its relationship to the present resource. Corresponding term list: Is part of, Has part, Is version of, Has versions, References, Is referenced by, Is based on, Is basis of.

- **NIEHS-DC Metadata Element: Format:** The physical or digital manifestation of the resource (e.g., HTML, SGML, XML, JPEG, GIF, etc.). Corresponding term list: text/html, text/rtf, text/pdf, image/jpeg, image/gif, image/tif, video/mpeg, video/quicktime, audio.

The term lists the team incorporated into the NIEHS-DC metadata schema were designed to assist the metadata creator in selecting appropriate descriptive terms from a finite set. In addition to Format, Relation and Type above, term lists were adapted for Language and Description. The terms chosen for these value lists were compiled with close attention to those already accepted by the Dublin Core Metadata Initiative for element qualifiers (http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/) and the DCMI Type Vocabulary (http://dublincore.org/documents/2000/07/11/dcmi-type-vocabulary/). The list for the NIEHS-DC Language element is limited to English, Spanish, French, Chinese, Japanese, German, and Other. At this time, the languages of the NIEHS Web pages can be adequately described from this list. The team created a term list for the NIEHS-DC Description element to include Article, Brochure, Bibliography, and Directory. In addition, the team developed a term list for the locally defined NIEHS-DC Audience element that includes the following terms: Researchers, Kids, Students, Teachers, General Public, NIEHS Employees. Since many of the NIEHS Web pages are designed to target specific groups of public users, it was important to provide accurate vocabulary choices to distinguish between various kinds of audience groups. The team also defined a default or fixed value of "NIEHS" for the

https://doi.org/10.23106/dcmi.952106581

NIEHS-DC Publisher element and a default value for the NIEHS-DC Rights element stating that NIEHS is a Government agency and U.S. Government works are generally not eligible for copyright.

5. The choice of XML

Based upon the experience of several years of change in Web practices and computing trends, the NIEHS metadata team decided to implement a metadata storage system that would endure further changes as much as possible. For a Web system, this meant utilizing standards-based, platform-independent and well-supported components wherever possible.

The chosen method of storage would need to facilitate:

1. Indexing and retrieval by the in-house Ultraseek search engine.
2. Indexing and retrieval by external WWW search engines such as Google, AltaVista, Northern Light, etc.
3. Immediate use of, or potential migration to, an Oracle (NIEHS standard) database.

In theory, XML seemed to be the best choice: a simple text-based, highly structured markup metalanguage, recommended for several years by the World Wide Web Consortium. XML held the promise of being transportable, flexible, and readable by both humans and computers. A non-proprietary XML data foundation could become the basis of numerous other valuable Web and library development projects in the coming years at NIEHS.

The NIEHS Ultraseek search engine allows XML Element-to-fieldname Mappings enabling the Dublin Core (or any other) metadata elements and attributes found within XML tags to be included in the search-and-retrieval process along with metadata derived from Ultraseek's regular document indexing. However, efforts to identify models for the NIEHS system based upon existing functional XML metadata implementations were unsuccessful. Inquiries with the technical support office at Ultraseek about the feasibility of developing a mixed collection of existing documents (HTML, PDF, etc.) and standalone metadata XML files resulted in the response that it had not been implemented previously by anyone. Considerable alterations to the source code would be necessary.

Therefore, it was critical to confirm that other key components of a proposed system would work. First, the team successfully tested the indexing and retrieval functions of the Ultraseek server on standalone XML metadata files. Next, they determined that since XML and HTML were both text-based markup languages, a script could be adapted or written (if necessary) that could grab the data from the XML files and insert it back into the original HTML documents as metatags. This would better accommodate the major WWW search engines, which do not currently spider Dublin Core element tags. The news media reported that Oracle intended to embrace XML, and then Oracle followed up with release of a host of XML tools. Assured that the NIEHS metadata team could (if necessary) move the metadata into an Oracle database, taking full advantage of well-known performance advantages, the team reconfirmed the plan to store the metadata in XML files initially.

All that was left to determine was how to optimize the utilization of the XML metadata files. The team weighed the advantages and disadvantages of three possible approaches:

1. Use of metadata XML files on a standalone basis.
2. Use of both XML files and original HTML documents enhanced with metadata extracted from our XML files.
3. Use of both XML files and an Oracle database containing data extracted from the XML files.

While solution number one has the advantage of providing more precise metadata (and search results), it also carries the disadvantages of problematic search engine source code alterations, limited or no availability to external WWW search engine spiders, and a confusing either/or choice for users of the in-house Ultraseek search. Solution number three may be more attractive in the future, but for now it also carries some burdensome source code alterations. The metadata team chose solution number two, which promises relatively immediate improvements to both users of the in-house search engine *and* users of external WWW search engines who search the NIEHS Webspace. Some programming will be required to extract metadata from the XML files and insert it in the form of metatags into the original documents, but this can be accomplished simply by adapting scripts previously written for this purpose.

Using the Dublin Core metatags relieves the problem of the perceived ownership of the Web pages by the content creators themselves. NIEHS has a tradition of allowing a high degree of independence in Web page creation, and care must be taken not to give the impression to the content creators that someone else is tampering with their pages. By integrating the metadata in distinct Dublin Core metatags, the content creators and Web developers can maintain ownership of their pages and the default HTML metatags. The DC metatags can simply be added to the headers of the pages and also reside in standalone XML files.

6. Conclusion and future work

https://doi.org/10.23106/dcmi.952106581

The NIEHS metadata team has completed the first step in enhancing access to the organization's Web pages by conceptualizing the NIEHS-Dublin Core metadata schema and developing it in an XML structure. The NIEHS metadata team has united all 15 elements from the Dublin Core namespace and a single element from the DC Education namespace to create its application profile. As part of this activity, several definitions were refined for public viewing and term lists were added, although the refined elements still fit the DC formal definitions. Future issues center around subject taxonomies, controlled-vocabulary versus free-form keywords, evaluation of the effectiveness of the customized elements, and the ability of content creators to create their own metadata. The team has already conducted a baseline study of the ability of content creators to create metadata[6], and members of the team will continue to examine this issue.

Development of the NIEHS-DC Metadata Schema has raised a number of policy issues with which an organization must deal. Should metadata be assigned centrally in an organization? If so, where in the organizational structure? Should the organization accept the metadata as provided by the content creator or should it give preference to metadata assigned by a professional metadata creator (cataloger)? How does a diversified research organization reach a compromise between the independence of its investigators and the need to speak with one corporate voice? How do you get the content creators to "buy in" to the need for metadata? How do you get them to use a prescribed form and stick to a standard? How do you ensure that new metadata is generated when changes and updates are made to a Web page? How do you decide which content creator is responsible for the metadata when more than one person contributes to a Web page? These are the issues the NIEHS metadata team is actively pursuing.

Acknowledgement

We would like to thank Cristina Pattuelli, Bijan Parsia, Zhihui Liu, and Eun Hyung Doh of the University of North Carolina at Chapel Hill, supported by Microsoft Corp., for their help in using XML and in developing the schema.

References

[1] S. Weibel. Metadata: the Foundations of Resource Description. *D-Lib Magazine*, July 1995. Available: http://www.dlib.org/dlib/July95/07weibel.html.

[2] S. A. Sutton. Conceptual Design and Deployment of a Metadata Framework for Educational Resources on the Internet. *Journal of the American Society for Information Science*, 50(13):1182-1192, 1999.

[3] P. Miller. *Metadata for the Masses.* Available: http://www.ariadne.ac.uk/issues/metadata-masses.

[4] Weibel, 1995.

[5] S. J. Darmoni et al. The Use of Dublin Core Metadata in a Structured Health Resource Guide on the Internet. *Bulletin of the Medical Library Association*, 89(3): 297-301, 2001.

[6] J. Greenberg et al. Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. (Draft, DC-2001 Conference).

https://doi.org/10.23106/dcmi.952106581