The Dublin Core Descriptive Metadata Program: Strategic Implications for Libraries and Networked Information Access

by Clifford Lynch, Executive Director, Coalition for Networked Information

This article explores the potential applications of the Dublin Core descriptive metadata program for libraries, museums, and other networked information providers. It includes a brief summary of the current thinking that has emerged from the Dublin Core initiative, including the broader metadata framework captured by the Warwick Framework, and also some consideration of the relationship with the Resource Description Framework under development by the World Wide Web Consortium, but the central focus is on applications scenarios for Dublin Core metadata. The approach here is strategic rather than technical; readers interested in the precise definitions of the individual Dublin Core data elements will need to consult the actual Dublin Core documents or other articles. My hope is that it will be helpful to library administrators and technology managers who are trying to understand and evaluate the implications of the Dublin Core both for access to existing resources and for practices of description that will be used to create, maintain and provide access to new resources.

Metadata

Metadata is literally "data about data," information that qualifies other information. Bibliographic description is a form of metadata, so also is information about intellectual property rights and terms of use, formats of electronic information, reviews, errata, abstracts and summaries, provenance information, and a host of other data. Some metadata can be derived mechanically from objects; other metadata has independent standing as intellectual creation in its own right. It should be clear that the set of metadata associated with an information object is unbounded. The division between data and metadata is somewhat arbitrary and highly situational; information will be used as data in one setting and metadata in another.

At least in my view, discussions of metadata independent of context and purpose are of little interest; it is most productive to speak of various kinds of metadata in conjunction with the processes that they are intended to support or facilitate. There are certainly types of metadata that have been developed for various specific purposes and which it is now proving possible to repurpose, particularly in the digital environment—indeed, creative repurposing and reuse of metadata is emerging as a key idea in the development of sophisticated information organization, retrieval, and management systems. But the point is that metadata is created and takes on importance through its ability to support activities; for example, the point is not to describe but to support discovery and other processes.

For a more extended discussion of the nature of metadata and its interactions with the processes of the discovery and retrieval of networked information, readers might consult the draft CNI white paper on the topic http://www.cni.org/projects/nidr/.

The Warwick Framework

While the Warwick Framework (named after the meeting in Warwick, England where it was developed) actually postdates the beginning of work on the Dublin Core (DC) described later, it is useful to discuss it first because it provides a broad framework in which to define sets of metadata.

The basic motivation for the work on the Dublin Core was to develop a set of simple data elements that could be used to describe document-like networked information objects in support of discovery (searching) activities. It rapidly became clear that there were any number of legitimate, important requirements for types of metadata that went beyond the scope of the Dublin Core; the problem was that because the Dublin Core was an active effort, and also because it was not clear how to use the DC in conjunction with other sets of metadata, there was considerable pressure to extend the scope of the actual DC effort almost without boundaries. This threatened the effectiveness of the Dublin Core program. To address this problem, an architecture called the Warwick Framework was developed that described how various sets of metadata for different purposes might be defined and maintained by appropriate communities of expertise. Collections of data elements from these diverse sets of metadata would be assembled into "packages" (one package per metadata set). The framework describes container structures whereby a digital object and a collection of such packages can be linked together. Each package is independent of all of the others, and software systems that understand specific metadata sets can extract packages that are based on those sets and examine them, bypassing other packages based on unfamiliar sets. Individual packages can even be encrypted independently. Containers can also refer to remote packages stored independently on the network, and are recursive: a container can include other containers, allowing for the construction of complex composite objects.

In designing the Warwick Framework, there was a recognition that division of the universe of metadata into packages would be imperfect; there would be some overlap between packages, and the content of one package might, in some cases, be derived computationally from another. There are also a number of research questions about how relationships among packages are expressed.

The importance of the Warwick Framework is twofold. First, it provides a broad architectural framework for defining and using metadata of various types. Second, it allows developers of metadata sets that have specific purposes to limit and focus their work by appealing to the Warwick Framework as an overarching context within which other groups interested in metadata can independently make progress on their own needs.

More information on the Warwick Framework can be found in the article by Carl Lagoze in the July 1996 issue of *D-Lib Magazine* http://www.dlib.org/dlib/july96/lagoze/07lagoze.html and the references there.

The Dublin Core

The Dublin Core is a set of fifteen data elements—each of which is both optional and repeatable—that was designed to be used as metadata to describe a broad class of information objects. The description applied to objects through the Dublin Core data elements is not

intended to be comprehensive or exhaustive; it does not seek to capture everything that can be asserted about an object. In particular, the DC is designed to support discovery of information objects of interest using search tools and systems; it is not designed to provide comprehensive support for access, management, use, or assessment of networked information (though some of the metadata to support discovery is also important in these contexts). To give just one example of this distinction, the Dublin Core includes a data element for terms and conditions. This is provided primarily because some limited information on terms and conditions of use—for example, that an object is not copyrighted, or that it can be used without restriction in educational settings—is actually important in finding objects of interest. There is work underway to develop very complex codings to express terms and conditions that might be used in conjunction with electronic rights management systems; this data would be essential for use and management applications, but is probably too detailed and specialized to be of much use in the discovery process, particularly given the current immature state of both standards and conceptual understanding in rights management specifications. The Dublin Core is not intended to carry this kind of very specific functional rights management terms and conditions metadata.

The DC was developed to describe what have been called, for want of a better term, *document-like objects*. These have the characteristics of being relatively fixed, although they need not be textual (images or sound recordings are easily within scope). They may have internal sub-structure—for example, an object with component objects—but the main focus of the DC to date has been to describe objects as opposed to *collections* of objects. The primary concern has been to ensure that the DC is serviceable for a rather broad range of common information objects—for example, a workshop was held with the specific focus and outcome of extending and validating the DC as a means of describing a large class of visual resources—rather than drawing precise boundaries for what is out of scope.

The DC clearly *can* be usefully applied to collections or to very complex dynamic objects or information services, but it probably does not do a completely adequate job of describing such objects and services to support discovery.

The development of the DC has had a very strong theme of codifying practice rather than research: methods for satisfactorily describing these complex new classes of dynamic digital objects and network services is still, at least in my view, a research problem. Further, while there is a great deal of consistency across the class of document-like objects that were the objective of the DC, there seems to be tremendous variation in the kinds of description needed for the still evolving menagerie of new digital resources. And, at least today, there are a lot more document-like objects than anything else on the network; they are the rule rather than the exception.

Data Elements

The fifteen Dublin Core data elements are: title; creator (author); subject and keywords; description; publisher; other contributer; date; resource type; format; resource identifier; source; language; relation; coverage; and rights management.

It should be clear from an inspection of this list that the DC is designed to serve as a sort of lowest common denominator form of description. It does not, at least directly, accommodate discipline-specialized description; indeed, some of the data elements, such as "date" are so vague that they are of limited utility without some further scoping. The DC was designed so that data element values for an object could reasonably be defined by its author, or by a site manager, rather than by a trained specialist cataloger or indexer. DC relies very heavily on natural language, and retrieval systems for the DC will have to rely heavily on heuristics and language parsing; not only does the basic DC framework ignore specialized thesauri and subject classification, for example, but it does not even make assumptions about the format that would be used to list dates or personal names.

Qualifiers

One of the ongoing tensions and controversies in the development of the DC has been its lack of precision. The most basic version of the DC—called the *Unqualified* Dublin Core—doesn't carry any information about the format of the data element values, their source or context, or the specifics of their semantics beyond the very broad definitions of the basic data elements. To address this need various qualifiers have been proposed to serve such functions as indicating the language or syntax in which the data element values are expressed, or to constrain the semantics of the data elements (for example, indicating that a date is a date of creation, or that a creator is a corporate author, or that a topic value is taken from a specific thesaurus). Obviously, use of qualifiers will tend to reduce interoperability, because participating systems will need to understand much more than just the fifteen basic data elements in order to interpret the semantics properly. To address this problem, a basic rule has been established for all types of qualifiers: if one ignores the qualifier, the data element value must be consistent with the basic definition of the data element's semantics in the DC. Thus, qualifiers can only constrain or refine the semantics of the DC data elements; they cannot be used to alter their meanings so that they are inconsistent with the original definitions. Definition of the data elements which should under normal practice be qualified, and what the appropriate values of these qualifiers should be, is a subject of ongoing work within the DC community; in a sense, this can be viewed as a discussion about how to extend the DC beyond the original fifteen elements in practice without destabilizing the original definitions, although some qualification (for example, to indicate the language or format of the data element value rather than the meaning of the data element itself) really has a different and less semantically significant character. It remains to be seen how the use of qualifiers will evolve within the various communities of DC users.

Relationship to Surrogates

Another controversial issue in the definition of the Dublin Core has been how use of the DC elements should interact with the surrogates that are so commonplace in the digital environment. For a document that was created as a digital object, matters are simple: the DC data elements describe the document. The creator of the document is the person who authored it. But consider this common case: there is a painting hanging in a museum that was created by artist X; fifteen years ago, photographer Y took a picture of this painting; last week, curator Z digitized Y's picture of X's painting. What are the semantics of the DC metadata associated

with the digitized image? The answer is that there should be three groups of DC metadata: one for the painting, one for the photograph, and one for the digitized image. The first would list X as creator; the second Y, and the third Z. The three groups of DC metadata would be connected through the "relation" data element. This has the advantage of being conceptually simple, albeit a bit verbose, for those creating metadata (though this can clearly be mitigated by a well-designed data entry system for DC metadata). It also places a considerable burden on the design of retrieval systems to behave intelligently: to many users, the conceptual distinction between painting, photo, and digitized image is at best murky, and an end-user query will often ask for the painting when what the user really wants is a digitized image of the painting. Retrieval systems will need to be able to retrieve clusters of groups of DC data elements and present them to the user in a comprehensible fashion.

The Evolution and Documentation of the Dublin Core

To date, the DC has been developed informally by a loose international consortium of interested parties through a series of five workshops: Dublin, Ohio (from which the core takes its name); Warwick, England; Dublin, Ohio again (a meeting focused specifically on the role of the DC in describing visual resources); Canberra, Australia; and Helsinki, Finland. The sponsors of these meetings have included OCLC, the National Center for Supercomputing Applications, the National Science Foundation and the Coalition for Networked Information in the U.S., UKOLN in the U.K., the Australian National Library, the National Library of Finland, and many others. Stuart Weibel of OCLC has been the leader of the effort since the beginning.

At the conclusion of the Helsinki meeting in late 1997, a series of working groups were chartered to continue efforts to extend and refine various aspects of the DC. It is likely that work in 1998 will proceed through a series of smaller meetings focused on specific issues, concluding with a sixth plenary meeting late in 1998.

At present, the Dublin Core is documented in a series of meeting reports and articles in *D-Lib Magazine*, and in working documents on the DC website http://purl.oclc.org/dc/; this site includes extensive information on the meetings, bibliographies, and other useful links. A series of informational (not standards-track) IETF RFCs are in preparation and should be released within the next few months. There are ongoing discussions about progressing the DC through the U.S. National Information Standards Organization as a formal standard, and also about what should be done to provide an ongoing "home" and maintenance agency for the standard if and when it is finalized.

Note should also be made of the work of the World Wide Web Consortium, which is working on a program they call the Resource Description Format (RDF). While this work is not directly driven by the DC, and in fact has some of its roots in extending earlier Consortium efforts to develop PICS (the Platform for Internet Content Selection) for rating and content filtering applications, the group working on RDF includes heavy representation from the DC community. The goals of the RDF effort include the definition of general mechanisms for attaching metadata of all kinds to web pages composed using the new Extended Markup Language (XML) defined by the Consortium, including DC metadata, the development of

schema definitions for metadata sets, and query facilities for metadata. This work will likely be central to facilitating the large scale use of DC within the Web.

More information on PICS, RDF and XML can be found at the Consortium's website http://www.w3c.org/ .

Machinery Needed to Support the Use of the DC

At one level, the Dublin Core is a conceptual construct; it captures the idea that there are pieces of text that can be associated with an information object with agreed-upon semantics such as those of "creator" or "relation." In order to make this conceptual construct concrete and to apply it in the networked information environment—which is characterized by the cooperation of large numbers of autonomous machines and agencies, and the sharing of information among them—there is need for a variety of supporting machinery. This machinery is codified in supporting standards and practices. It's important to recognize that, in a sense, the Dublin Core transcends specific machinery, and that many different mechanisms can legitimately be developed within different communities of practice and implementation to meet these requirements. It is likely that over time new mechanisms will continue to develop as a result of the overall evolution of architectures and standards for the networked information environment.

The three major classes of mechanisms are: encoding and transfer syntaxes; methods of associating or attaching groups of Dublin Core data elements with the information objects that they describe; and, more generally, methods of retrieving or querying Dublin Core data associated with objects or groups of objects.

Several methods have been proposed for encoding groups of DC data elements for storage and inter-system exchange: these include the use of HTML META tags in today's HTML-based web pages; the use of XML structures as specified by the World Wide Web Consortium's Resource Description Framework (RDF) as part of future XML-based web pages; and the incorporation in SGML. Several of these proposals also address the problem of associating DC elements with objects in a very direct way: in the Web setting, they are simply incorporated as part of web pages. There is also a way of requesting DC information for an object via HTTP (thus leaving it up to the web server to maintain the linkage between DC elements and the base object internally); this mechanism can also be used to query third-party metadata servers for DC metadata.

One of the key deployment scenarios envisioned for the Dublin Core is that web pages will increasingly incorporate DC data elements as part of the pages—using either direct coding in META tags for current HTML pages or the new RDF structures for pages in the newly defined XML format—and that the familiar web indexing programs (or their successors) will be upgraded to capture this metadata and incorporate it into their web indexes, so that one could query a system like Lycos or Alta Vista for pages that have a specific creator, for example. This metadata might be created by the authors of the pages, by website managers, or by third party indexers/catalogers. Complementing this, we are likely to see third party databases of DC metadata develop which simply refer to and describe web content and other information

objects.

It's essential to recognize that while the Web—and in particular the static, visible web of HTML pages—is a key applications environment for DC, it is not the only one. It is perfectly reasonable to think in terms of databases containing objects described by DC data elements; here the DC data elements would be encoded and linked through some local data structures. The retrieval of an object from such a database—accomplished through an interactive forms-oriented query interface or an inter-system query protocol like Z39.50—might cause the retrieved object to be encoded as a well-known, common format, such as a page that included XML tags for the relevant DC elements. Similarly, one might want to associate DC elements with an entire website or database; here one would need a mechanism (perhaps akin to some of those used in the Harvest system) that could be used by network resource indexing systems that build site or database directories. (For more on these issues, see Clifford A. Lynch's "Searching the Internet," *Scientific American* 276.3 [March 1997]: pp. 52-56, available at http://www.sciam.com/0397issue/0397lynch.html.)

At present, query facilities for Dublin Core data elements are very diverse. There are a number of interactive query systems that offer DC data elements as access points to specific databases or other information collections. Several Z39.50 attribute sets—notably GILS and BIB-1—are incorporating the DC elements as access points that can be used in query construction, and, as part of the migration to the new Z39.50 attribute architecture, it is likely that a separate Z39.50 attribute set will be defined. Part of the RDF work program includes the definition of query facilities for metadata; however, work on this is only at the earliest stages.

Applications Scenarios for Libraries

The Dublin Core has two different basic applications for libraries. The first is in permitting library databases to become part of broader network search services, or to allow libraries to provide their patrons with consistent views of both library and non-library databases. The second is in describing new resources that cannot be cost-effectively supported through traditional cataloging approaches.

Use of DC in Federating Existing Resources

One of the key notions in networked information discovery and retrieval is that of *federating* disparate, independently maintained databases scattered about the network. Users should be able to search such constellations of databases as if they were a single, consistent, unified information resource. In order to do that, it is necessary to provide a common semantic view of the various databases involved, even though they may have radically different access points and data structures, and may be accessed through different search protocols or other query mechanisms.

Because the Dublin Core is designed as a lowest common denominator descriptive approach, it offers a very flexible and general context to support federation. Traditional library catalogs or abstracting and indexing databases can clearly support queries constructed using Dublin Core data elements (albeit with some reduction in the precision that queries can express as

compared to queries formulated using the database's native search language, unless qualifiers are used extensively); thus it is possible to build a software layer that permits such databases to participate in federations that use the Dublin Core data elements. These interfaces will use mappings or crosswalks to translate from Dublin Core data elements to the actual access points in the database. Mappings have already been developed from DC to MARC fields.

I think it is likely that libraries will use this capability to make their databases visible in database federations that operate outside of the traditional library systems and services—these databases could be searched as part of a distributed search system that also encompassed web-based resources outside of the library. To illustrate, suppose that a group of art history scholars developed a database of digitized images that were described using DC, and build a search system for that database. By making a DC-based "view" of a resource like a library catalogue or an art history abstracting and indexing database available on the Net, it would be possible to easily extend their system to also consistently search across the database and catalogue as supplemental resources. Or a system designed to search digital instructional media might be extended to also search library holdings through the same interface.

Conversely, because the Dublin Core is applicable to so many information resources, a library might develop a search interface and distributed search service that offered patrons a federated view of a very diverse set of databases, including not only traditional library databases, but also databases from other sources, such as government databases, databases produced by the next generation of web indexing services, or special purpose scholarly databases. While such a search service would not eliminate the need for much more precise and capable domain-specific and database-specific search facilities, it would be very useful to some users both in identifying databases of interest which they might then search directly for more comprehensive and precise results, or in doing very broad (not but necessarily precise or exhaustive) searches across a wide range of resources. In this connection, it is interesting to note that the Instructional Management System (IMS) being developed by Educom's National Learning Infrastructure Initiative (NLII) is using a descriptive scheme based in part on the Dublin Core for instructional media; this is a good example of a possible new resource that libraries may want to bring under the umbrella of a search system that also covers their catalogue and abstracting and indexing databases. (For more details on the NLII and its IMS project, see http://www.imsproject.org.)

Use of the Dublin Core in Describing New Content

The Dublin Core—perhaps supplemented by additional metadata packages defined within the Warwick Framework—will be used to describe content where traditional cataloging approaches are too costly, or where there is a need to create metadata for content that is not well served by current cataloging practices. The NLII IMS is a good example: many of the key things that users need to know in searching for instructional media can only be captured by traditional cataloging in unstructured textual notes. The IMS supplements the DC elements with an additional descriptive package designed specifically for instructional media. For digitized images or other materials, whether created directly in digital form or digitized from other media (e.g., special collections), full bibliographic cataloging is particularly expensive because most of these items are unique, and libraries cannot use the system of shared copy

cataloging to control and distribute costs. It's important to note that, while the Dublin Core was designed to be simple and thus much less expensive to apply than traditional AACR-2 based original cataloging, there is relatively little experience with it, particularly when the DC is supplemented with additional metadata packages. One effort that needs to take place over the next few years as part of the experience in using Dublin Core is some measurement of the cost savings over traditional cataloging for various types of material. We also will need to understand how retrieval quality varies with the different descriptive approaches.

The use of the DC and the Warwick Framework gives libraries the ability to design supplementary metadata sets —descriptive and otherwise—to characterize materials that either require more depth or precision of description than the Dublin Core alone can offer, or need not only descriptive but also other types of metadata associated with them in order to support processes that go beyond discovery (e.g., management, use and reuse, or rights clearance). Instructional media objects are a good example of such a category of materials; statistical datasets are another. Museums will likely make substantial use of the DC plus additional metadata packages. To a great extent, I suspect that library use of the DC for description will be determined by the policy choices that libraries make about their role in creating descriptions for materials that have not historically been part of the mainstream of library collections, as opposed to simply making use of descriptions for these materials created by other (non-library) organizations.

Conclusions

The Dublin Core is clearly, in my view, going to be important for libraries both as an engineering tool for federating library and non-library databases, and also as a lower-cost alternative for describing materials. The creation of Dublin Core descriptions is going to be of particular interest for libraries expanding their collections with large amounts of digital content: images, sound recordings, video recordings, and new genres of networked information. In the longer run, I think it will also be important for libraries to track the work on the implementation of the Warwick Framework and to monitor the definition of additional metadata sets within that framework, which will be needed to address issues such as provenance, integrity, and management of digital content.

Editor's Note: NINCH, the National Initiative for a Networked Cultural Heritage, is the U.S. distributor of a 1997 U.K. report *Discovering Online Resources Across the Humanities: A Practical Implementation of the Dublin Core.* Contact ARL Publications for order information pubs@arl.org.

Table of Contents for Issue 196 | Other Current Issues Articles | Other Networked Information Articles


ARL Newsletter Home ARL Home

Maintained by: ARL Web Administrator Site Design Consultant: Chris Webster **Last Modified:** August 4, 2002