
Casting the Net

Caplan, Priscilla. "You Call It Corn, We Call It Syntax-Independent Metadata for Document-Like Objects." The Public-Access Computer Systems Review 6, no. 4 (1995): 19-23.

A few months ago I participated in the OCLC/NCSA Metadata Workshop. The name is perhaps a bit confusing, but the idea behind it was simple: we have to start defining a simple, usable standard for describing network-accessible information resources. The plan of action was also simple: find a group of interested people from a variety of constituencies, make them come to Dublin, Ohio, and don't let them leave until they've agreed on a core data element set.

The official record of the workshop is the OCLC/NCSA Metadata Workshop Report by Stuart Weibel, Jean Godby, Eric Miller, and Ron Daniel. I'll summarize briefly here with a little bit of the context, but if you're interested in more detail, the full report is available at http://www.oclc.org:5046/oclc/research/conferences/metadata/dublin_core_report.html.

What is Metadata, Anyway?

Metadata really is nothing more than data about data; a catalog record is metadata; so is a TEI header, or any other form of description. We could call it cataloging, but for some people that term carries excess baggage, like Anglo-American Cataloging Rules and USMARC. So to some extent this is a "you call it corn, we call it maize" situation, but metadata is a good neutral term that covers all the bases.

The fact is, we already have any number of standards defining metadata element sets, from AACR2 to GILS. It is also a fact (and perhaps more important) that the vast majority of resources available on the network have no metadata associated with them at all. The goal of the workshop was to define a set of data elements simple enough for authors and publishers to use in describing their own documents as they put them on the Net, but useful enough to facilitate discovery and retrieval of these documents by others. This simple metadata could then also be used by catalogers and other third parties as a starting point for creating more rigorous or more detailed descriptions.

Our mental mission was to try to define a single sheet of guidelines that could be handed out to just about anyone. This had a lot of implications: it meant the core element set had to be short, the data elements had to be reasonably easy to understand and to provide, and the elements had to apply broadly to most of the resources under consideration.

The Dublin Core

Deciding just what resources were under consideration was an interesting issue. We finally limited ourselves to "document-like objects" or "DLOs," but we didn't waste much time defining what a DLO was, or what it wasn't. To me, an electronic text, map, or image would clearly be a DLO. To some participants, only textual materials qualified, while to others, computer systems or even people could be DLOs. The important thing was that the core data element set did not have to handle every type of resource that could theoretically be available on or through the network. It had only to handle things like, well . . . document-like objects.

The core element set itself, sometimes referred to as the Dublin Core, is still a moving target. The standard proposed in the workshop report will change over time as more people work on it and with it. As I write this, it consists of 13 elements. The first five elements, author, title, subject, publisher and date of publication, are for author, title, subject, and . . . you get the idea. The remaining elements are summarized in the report as follows:

> OtherAgent: The person(s), such as editors and transcribers, who have made other significant intellectual contributions to the work.
>
> Identifier: String or number used to uniquely identify the object.
>
> 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.
>
> Relation: Relationship to other objects.
>
> Language: Language of the intellectual content.

> Source: Objects, either print or electronic, from which this object is derived, if applicable.
>
> Coverage: The spatial locations and temporal durations characteristic of the object.

Every data element is optional and repeatable. Each element can also have subelements to qualify or explain the content of the element. For example, the subelement called "scheme" can be used to identify the authority or rule set used to provide the data element. Subject could have a subelement called "scheme=LCSH" to identify an LC subject heading, or "scheme=abstract" if an abstract were provided instead of subject headings.

You may be wondering how it could have taken 50 people three days to come up with a list of core elements. But the issues are fairly complex. Consider the concept of Author, for example. Earlier drafts had a single element (ResponsibleAgent) for all parties contributing to the intellectual content of the work, very much like a 7xx added entry field in USMARC. The type of agent, like author or illustrator, would be identified in a

subelement called "role." But some people argued that the concept of Author is so intuitive it needed to be singled out. If that's the case, what must be known about Author? In USMARC you need to indicate whether the author is a person, conference, or company; determine if the author should be a main or added entry; and decide what type or form of name it is. For the core element Author, you might be able to make these distinctions with qualifying subelements, but we weren't willing to be more restrictive than to require that names be in inverted order or sort order. The real wonder isn't what took us so long, but that we came up with anything at all.

One of the most important things to note about the proposed metadata standard is that it doesn't prescribe how to record this information; it is "syntax independent." The data may be represented in USMARC, HTML, SGML, or "keyword=value" pairs. Whatever you can process and exchange with your community is fine. This is a data element set, not a transport syntax. If you want to add data elements to the list, that's fine too. If you're part of a special community working with a special type of object, you can even define domain-dependent extensions, just as long as you don't expect everyone else to understand them.

What gives us hope that the Dublin Core will ever be adopted? I think it's partly a function of need and partly a function of the people who participated in the workshop. At times, the meeting reminded me of the bar scene in "Deep Space Nine": dozens of alien species milling about and talking slightly different English-like languages. There were the librarians, mostly middle-aged women like me, enduring endless jokes about sensible shoes. There were the IETF guys, astonishingly young and looking as if they were missing a fraternity party to be there. There were TEI people, GILS people, and geospacial metadata people. There were publishers and software developers and researchers. Every participant had his or her own perspective, and a few people had axes to grind, but nearly everyone agreed that there was a tremendous need for some standard and that an imperfect standard would be better than no standard at all.

Of course, there is a danger here: it's easy for anything to seem better than nothing until you have it. Non-librarians working on the metadata standard need to acknowledge that there are reasons why cataloging isn't simple or easy and that, with every simplification you make, you lose something in consistency or capability. Librarians need to see this initiative as a complementary, not a competitive, effort that needs our participation to succeed. This is an area full of tradeoffs, and I don't believe we've yet found the appropriate balance between effort and utility. We will, though, if we keep working at it.

Now that the workshop is behind us, the real work begins. A small group of participants have drafted the workshop report and will continue to refine the data element set. Those of us involved with particular data formats like USMARC or SGML must work in our respective communities to see how these metadata elements can be represented in our own syntaxes. Librarians have already begun their efforts with two discussion papers on the

Dublin Core presented at MARBI during ALA in Chicago this summer: "DP86, Mapping the Dublin Core Metadata Elements to USMARC" and "DP88, Defining a Generic Author Field in USMARC." Like the librarians, workshop participants involved in standards like the URC (Universal Resource Characteristics) need to investigate how the metadata standard fits into their existing standards. Those who are writing software need to see how they can incorporate metadata into their software.

Meanwhile, I'm going to sit back and meta-tate on how nice it will be someday to actually be able to find some of those document-like objects out there when I want them.

Appendix A: Metadata for This Article

Author = Caplan, Priscilla
Title = You Call It Corn, We Call It Syntax-Independent Metadata for Document-Like Objects
Publisher = The Public-Access Computer Systems Review
ObjectType = Article
Subject = Metadata
Subject = OCLC/NCSA Metadata Workshop
Subject = Cataloging Internet Resources
Date = 1995

About the Author

Priscilla Caplan, Assistant Director for Library Systems, University of Chicago Library, 1100 E. 57th Street Chicago, IL 60637. Internet: p-caplan@uchicago.edu.

About the Journal

The World-Wide Web home page for The Public-Access Computer Systems Review provides detailed information about the journal and access to all article files:

http://info.lib.uh.edu/pacsrev.html

Copyright