Catalogers Group

Minutes

October 11, 2001

Present: Angela Riggio, Beth Feinberg, Caroline Miller, Elie Chammou, Gia Aivazian, Janice Matthiesen, Jean Rashedi, Jeffrey Morehead, Joan LoPear, John Riemer, Luiz Mendes, Nancy Norris, Rebecca Aiken, Rita Stumps, Valerie Bross

Dublin Core Presentation (part 1)

John Riemer

Introduction

"Not everything can get a MARC record." I remember being really stunned by those words from Stu Weibel when I first heard them at a May 1998 ALCTS/LITA Metadata Institute in Washington. There are simply too many resources on the Web for each of them to receive a MARC record. The statement still applies when restricting our focus to just the worthwhile resources on the web.

Dublin Core (DC) data elements can be crosswalked to MARC—to what extent I plan on demonstrating.

What is *metadata*? (Structured) data about metadata. MARC is metadata, as much as Dublin Core is. While there are many metadata schema to pick from for a Catalogers Group presentation. I chose DC since it currently enjoys such wide, international usage, much like LCSH does in the realm of subject heading systems.

MARC is a very rich, highly-detailed metadata standard, with a long history and wide support. MARC is primarily used in libraries.

DC originally was intended for use by authors of web pages, to apply to documents they were creating. DC never caught on in that setting. As a tool for others to organize e-resources, it has come into favor in nonlibrary settings—more so than any other standard.

In comparison to MARC 21's hundreds of data elements, each with indicators and subfields, DC's 15 basic data elements look pretty "spare."

For e-resources, use of DC is a large step up from the specificity possible in search engines like Google. The search engines cannot distinguish among a name that is the author of the document, the subject of the document, or a mere passing mention within the document.

The actor Will Smith's name is made up of two rather common words. I understand there are a number of sites devoted to this actor, but only one official one issued *by* the actor < http://www.willsmith.net/intro_flash.html >. If the web site utilized the DC element for author in its header, finding the right web page could be quite easy.

As you might be aware, when DC was first proposed, some wondered whether it would become a threat to displace MARC. To say that not everything can receive a MARC record is not equivalent to saying it's a waste of time or overkill to use MARC for organizing *any* e-resources. From personal experience in using DC since 1999, I have had no regrets. I don't really see a risk of a dumbing down of cataloging to some new lowest common denominator.

Some more perspectives I have developed:

We already have multiple levels of records in the cataloger's repertoire. This would be just one more. We can put it to use as situations, staffing levels, and priorities warrant.

As you may know, DC records created in the CORC database are equated to Encoding Level 3; this sounds lower than Minimal-Level Cataloging (Encoding level 5 or 7), but DC records can and often do contain an indication of subject matter.

So long as e-resources are desired in the OPAC, to support one-stop shopping convenience, there will be a premium on the MARC format, on compatibility with the other records, and on comparable ability to retrieve across all types of material.

Records created in DC are not necessarily in their final resting state. Owing to crosswalks, any Dublin Core record can be mapped as-is to the MARC format. (You remember how DC & MARC are two views of the same record in CORC.)

When you mention DC records and MARC records, the tendency probably is to think of DC records as rather lean and MARC records as quite full. However, it doesn't have to be that way. Nothing prevents a DC record from being quite full; you have seen MARC records from vendors for belles lettres that are quite short.

I've been talking about how DC is not a bad thing; from a cataloger perspective, there ways DC looks like a good thing.

DC is a standard that can be used to create records for materials that would otherwise not get *any* cataloging. (Special Collections item-level descriptions for individual photos or letters are examples).

DC opens up the cataloging process to a lot more people. Some museums interested in sharing their resources on the Web may have no MLS types to perform or even guide the cataloging efforts. Special collections departments may have "Friends of the Library" volunteers who could participate.

Within libraries, which do have catalogers, we can similarly extend the labor pool for bibliographic control efforts to include reference librarians and selectors. In this setting you might to envision cataloging as an assembly-line process. They can bring information they know about the resource to the cataloging record. We would pick up where our collaborators leave off.

(For a long time catalogers have been welcome to serve on the reference desk. Until recently, it has not seemed like there was a tangible way for reference librarians to reciprocate. Now, by extending such things as BibCORC, we might well have a practical way to make the relationship a two-way street.)

For some resources receiving DC records, there may be no opportunity to further enhance the records. Still, the records can be crosswalked to MARC.

In other situations, we can look on DC records as being in a temporary state, that they are on their way to becoming what we are used to.

Since it is so unlikely we can obtain very many additional staff to match the appetite for e-resources in the OPAC, we probably really need to consider this strategy for extending the reach of our bibliographic control efforts.

So, how was it to first get involved in a DC project? Around Labor Day in 1999 I went to OCLC for a 2-½ day course for learning about and applying DC and practicing in the CORC database. Not everyone was a cataloger in the 10-student class, so we were highly encouraged to think and talk in DC the whole time, and it felt like one of those foreign-language immersion courses!

Soon afterwards I was approached about a organizing a cataloging project called Arts of the United States. In the 1950s a Georgia professor got a Carnegie grant to go around the U.S. photographing famous paintings, sculptures, decorative arts, and significant architecture. The negatives were used to create slide sets for two generations of art appreciation classes. They were about to be discarded by the last company that had any use for them, in New York. They were rescued and brought back to University of Georgia, which happened to hold the copyright on them. So, it was fair game to digitize them and put them up on the Web.

I was given a guidebook for the collection that gave cryptic information about each image, similar to the little placards you see next to art objects in a museum. The Art School was so excited at the prospect of being able to use the images on the Web, they offered from their own budget a bright grad assistant to do the cataloging in the library. I had about a week to design some procedures for her to use.

I'm distributing a copy of the latest version of that project's instructions. In the spirit of that OCLC workshop, I tried to maximize the amount of data entry done on the DC side, and requested MARC entry for only that part of the record that couldn't be done any other way. The grad assistant was fairly familiar with the Art and Architecture Thesaurus

(AAT), and since its 120,000 terms were easier to apply to the images, I decided to have her assign from that controlled vocabulary instead of LCSH.

Using the DC elements gave me an idea what it must have been like to use MARC when it was brand new. How to equate the MARC tags to what was paragraphed on printed cards was not well established at the beginning. Lots of procedural questions and ambiguities must have existed, as I faced in my project. I could see how valuable it was to have a MARC background, in that a non-cataloger would be unduly influenced by the connotation of the words that serve to name the various DC elements. As I was pigeon-holing the data from the guidebook into DC, I was also thinking ahead to how CORC would be crosswalking the data to MARC fields and to how well the resulting records would sit next to other MARC records, if we put them in the OPAC.

With 4500 records to be created, I felt a strong anxiety to think of everything at the beginning, so we would not have to revisit the records and make corrections. Recently a former UCLA Library employee Max Marmor came to campus to promote the art world equivalent to JSTOR (ArtSTOR), and I learned the company storing the images here in Venice was going to use another metadata standard called VRA Core and map from the DC records in CORC. Just about the only quibble they have with the CORC records is that the repository of the art object is not in a unique field that could be distinguished from all the other corporate added entries.

So, on to the 15 DC elements …

The Basic DC Element Set

The set of elements got its name from the site of the 1995 meeting that reached agreement on the first 13 of them—an OCLC-hosted meeting in Dublin, Ohio. According to the book, *Getting Mileage Out of Metadata*, they are seen as core elements that can be used in describing electronic resources. The elements use "easily understood names and single-word labels." DC is meant to be applicable in all disciplines and formats, while many other metadata schema are tailored to a particular discipline.

Use of data elements is meant to be highly flexible. All elements are optional, all are repeatable, all are displayable in any desired order. Occasionally you will see the 15 elements grouped into three major categories:

Content	**Intellectual Property**	**Instantiation**
Title	Creator	Date
Subject	Publisher	Format
Description	Contributor	Identifier
Type	Rights	Language

Source		
Relation		
Coverage		

Taken from: http://www.ietf.org/rfc/rfc2413.txt Sept. 1998.

In the handout, I've arranged them in alphabetical order since I thought that would make them easier for you to refer to in the future. The names are deliberately kept to single words, though several have an official longer name. Quoted commentary is taken from Dublin Core Metadata Element Set, version 1.1: Reference Description, July 2, 1999 http://dublincore.org/documents/dces/

Title: A name given to the resource.

Comment: "a name by which the resource is formally known." (Formerly: "The name given to the resource, usually by the Creator or Publisher.")

Subject [and Keywords]: The topic of the content of the resource.

Comment: This also includes classification. "Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme."

Description: An account of the content of the resource.

Comment: "Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.'

[Resource] **Type**: The nature or genre of the content of the resource.

Comment: "Recommended best practice is to select a value from a controlled vocabulary (for example, the working draft list of Dublin Core Types (10 general categories)

http://dublincore.org/documents/dcmi-type-vocabulary/

Collection Dataset Event Image Interactive Resource Service Software Sound Text

This element should be distinguished from the physical or digital manifestation of the resource, which is covered by the element Format.

Source: A Reference to a resource from which the present resource is derived.

Comment: "The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system."

Examples might be the work this resource was translated from (whole work) or excerpted from (subset)

Relation: A reference to a related resource.

Comment: "Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system."

Examples are earlier/later titles for a serial, part-whole relationships like the series the resource is an issue of or a building an architectural detail is a portion of.

Coverage: The extent or scope of the content of the resource.

Comment: Coverage is either spatial (geographic) or temporal (time period). Recently a third category, jurisdiction, has been added. Temporal coverage refers to intellectual content: what the resource is about. When the resource was created or made available goes under Date.

Creator: An entity primarily responsible for making the content of the resource.

Comment: This can be "a person, an organization, or a service."

Publisher: An entity responsible for making the resource available.

Comment: This can be "a person, an organization, or a service."

Contributor: An entity responsible for making contributions to the content of the resource.

Comment: This can be "a person, an organization, or a service."

Rights [Management]: Information about rights held in and over the resource.

Comment: This can contain either a rights management statement or a pointer to one.

Examples are "Property of …" or copyright statements, or other restrictions on usage.

Date: A date associated with an event in the life cycle of the resource.

Comment: Recommended best practice is ISO 8601:

http://www.w3.org/TR/NOTE-datetime

Example of a date going beyond a mere year: YYYY-MM-DD (eg 1997-07-16)

Format: The physical or digital manifestation of the resource.

Comment: May include either the media type or the dimensions of the resource.

Recommended best practice is following the Internet Media Types (MIME) list:

http://www.isi.edu/in-notes/iana/assignments/media-types/media-types

Examples: text/html, application/pdf, image/jpeg, video/mpeg, application/x-dejavu

[Resource] **Identifier**: An unambiguous reference to the resource within a given context.

Comment: "Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. " For electronic resource this obviously includes URLs. It could also include a system control number, e.g. an ISBN.

Language: A language of the intellectual content of the resource.

Comment: Recommended best practice currently happens to be ISO 639

http://www.oasis-open.org/cover/iso639a.html

These language codes are 2 characters long ('en' for English), and can take hyphenation of 2-character country codes for dialects ('en-uk' for British English)

So far, I have shown you the 15 basic data elements. They seem rather general, while the MARC 21 data elements are highly specific. It could be fairly difficult to automatically map DC to MARC under these circumstances. (Does anybody want to suggest how any of these would map to MARC tags?)

Equating Title to 245 and Publisher to 260 $b look like the most promising possibilities.

(What can we do with mapping Contributor to MARC? Probably the 720 field, if the system cannot discern the personal or corporate nature of the Contributor.)

Dublin Core Qualified (DCQ) Data Elements

The mapping to MARC, as well as the searching with DC elements themselves, can be greatly improved by the use of qualifiers. Qualified DC data elements are 3-tiered:

Basic element + Element refinement + Encoding scheme. Typically these are separated periods: Example: Subject.topical.LCSH.