10-14-1999

Cataloging Web Resources: CORC and Its Alternatives

Wendy C. Robertson
University of Iowa

Cataloging Web Resources: CORC and Its Alternatives

By Wendy Robertson
Presentation given at ILA 14 October 1999

When the University of Iowa Libraries first heard about OCLC's project CORC, we thought it would be the answer to all our prayers. It would give us a fast and easy way to deal with web pages, it would teach us Dublin Core in 6 easy steps, it would be intuitive to use, it would bring us cookies, wash our windows and baby sit our children. Needless to say, we were wrong. It didn't do any of those things. It has been very good for forcing us to think more carefully about how our library deals with web pages.

CORC is a research project. It became public in January 1999. We learned how to use it in March. It was still very rough, although it has improved quite a bit over the last 6 months. Time is too brief today to explain CORC in detail, however I have made a handout that introduces record creation in CORC. I will focus on our reactions to CORC and the direction in which our library now plans to go.

[CORC is intended to assist in creation of MARC or Dublin Core records, automatically harvesting some information to give a good start to the metadata creation. These records will then be available in a searchable shared database of records for web resources. CORC will translate records between MARC and Dublin Core. CORC will also be one central repository of records for web resources, so URLs only need to be updated in one place, and will hopefully be updated automatically.]

Harvesting information is the most exciting part of CORC from a cataloger's perspective. When I first used CORC in March, it became apparent that it translated between the web page, MARC and Dublin Core in odd ways. Instead of relying on CORC as an intermediary, I realized I needed to learn about Dublin Core to understand how CORC was behaving. *[I have also made a handout that gives a brief explanation to Dublin Core.]*

In my readings, I learned that Dublin Core is still a work-in-progress. It is a container for structured information; it does not provide content standards. These standards are still being developed. While the 15 elements seem stable for now, the qualifiers that CORC uses are not necessarily in use by the wider Dublin Core, or DC, community. Different organizations are allowed to make different qualifiers according to the needs of that organization. This is part of the flexibility of Dublin Core. Since a major feature of CORC is the ability to display records in either MARC or DC format, the CORC DC qualifiers tend to turn Dublin Core into MARC. DC was intended to be a fairly simple, flexible format, but CORC's version can have much of the complexity of MARC. Libraries already have MARC – I'm not sure why we need to have Dublin Core recreate it.

To add records to the CORC catalog, you can have non-catalogers learn how to create structured, standardized records, have catalogers learn Dublin Core, or have catalogers catalog in MARC format using CORC. Creating a MARC record in CORC would be beneficial if enough information was automatically harvested. I continue to be disappointed with the results of the harvesting. Although it has improved, CORC still has so many problems it doesn't really save me any time to have this content prompted. It is much faster to start from scratch or copy an existing record and paste information from the web page into the library catalog. I also found CORC to be extremely slow. I found I much prefer to create or update a record in the system I use everyday rather than spending additional time learning extra software.

Because CORC is still a research project, standards have not yet been worked out. Some libraries do very full cataloging, others do very minimal. There is no quality control, no assurance that a trained subject analyst has given it subject headings. A few libraries, such as Auburn University and the University of Minnesota, have their policies posted on the web[1] *[point to overhead]* and they are interesting to look at.

CORC also promotes itself as a search engine. In theory since all the pages have been selected and have been described, you can get much better hit rates than with any other search engine. The problem with

this is that CORC records are proprietary. While participants can get the records they create at the end of the project, no one really knows what OCLC will do with CORC. We are pretty sure that access won't be free. It troubles us that OCLC will be able to profit off of our cataloging. I suppose it isn't that different from the charges for searching WorldCat, but if you are to take full advantage of pathfinder features and URL updating, then you will leave your records in CORC and may need to pay every time you search your own records.

CORC's ability to harvest information is still intriguing. At this time, it cannot properly interpret frames or text presented as a graphic *[even when ALT tags are used appropriately]*. CORC harvests best from well designed text pages, particularly from pages that use the META tags "description" and "subject" which are widely used among trained web page designers. As more web pages contain metadata, harvesting will become more effective. As CORC is not the only software that will harvest metadata, we will continue to watch to see how this technology develops.

After we had used CORC for a few months we decided it wasn't the direction in which we wanted to go. We realized we faced 2 separate issues: how to deal with locally created web pages and how to deal with everything else.

Typically, Dublin Core is seen embedded in HTML, although a separate database such as CORC is perfectly feasible. To embed Dublin Core, the web page creator must put the DC into the page, thereby distributing metadata creation. There are a few search engines, although not many, which make use of Dublin Core tags. One reason there are few search engines that make use of the tags is that a very small percentage of all pages have Dublin Core in them. The more pages that use DC, the more search engines will use the tags.

Librarians are in a good position to promote embedded Dublin Core because DC gives a structured form for information about the web page and librarians understand such structures. Roy Tennant states:

"Metadata, simply put, is structured information about information. The key is "structured." In metadata as in cataloging, a free-text description usually won't suffice."[2]

In keeping with the Metadata Working Group's objective, we decided that our library ought to put DC tags into all the important Library web pages. After that, we will train web page designers, in the Library and in the University community, how to use Dublin Core. This project is moving very slowly both due to time limitations and also because we have not yet found adequate software. Ideally, the software will prompt values for fields with limited content possibilities and we will be able to customize the software with common information for our local community *[such as copyright information in the rights field]*. It must also have wide distribution rights so that web page designers can use it themselves. We are checking different Dublin Core tools listed on the DC page, but haven't yet found our ideal. Once Dublin Core is in our local pages, it will improve the search results in many search engines, not just in CORC.

The other issue that concerned us was access to selected web pages. Metadata usually refers to information about electronic resources, particularly web pages. This is one area Technical Services Librarians too often ignore, even though creating "metadata" for a resource is really "cataloging" it. CORC tries to support this cataloging, but it confuses Dublin Core with MARC and doesn't give any guidelines for content standards. As long as we are working with MARC, we wondered why we should go through CORC – why not simply add the records for web pages to our catalog?

Cataloging web pages can be a pretty horrifying idea. There are so many of them! How could we possibly catalog them all! We can't and we shouldn't. Your library doesn't buy every book in the world, does it? You wouldn't buy them even if you had the money (or the space). Libraries select books. They can also select web pages. If a web page meets standard selection criteria we should catalog it. In other words, if it was in print and you would buy it, then you are doing a disservice to your patrons to exclude it from the catalog simply because of its format.

Too many libraries are developing split personalities. Traditional resources are fully cataloged by people trained in creating (and who like) detailed, standardized, structured information. Reference librarians choose web resources and then make pathfinders to them. Pathfinders are great for the very best resources, but they quickly become unwieldy when they are several screens long. Multi-disciplinary resources need to be added to multiple pages and links need to be kept current in both places.

The University of Iowa has an improved version of the pathfinders. Our Gateway will search a database of selected web pages. The patron will then get a pathfinder on exactly the topic she wants. Our gateway database can be viewed as a miniature library catalog. Graduate assistants add items to it the best they can with no training in cataloging. Our catalogers have essentially passed the buck to untrained staff. In the best situations, we are still wasting effort; items already are in the library catalog so student workers can copy and paste many fields, improving the quality of information. Because it is a locally designed database, it also cannot take advantage of shared cataloging.

If we could instead put all the selected pages into our library catalog and link directly with the gateway, access to web pages would become much more straightforward. No longer would they be isolated as a special category. Web pages are part of daily life and ought to be incorporated fully into the library workflow. The University of Iowa has not yet migrated to a catalog with a web version. Our new system, Aleph, should be able to have a much more direct link between the catalog and the web, so that the gateway can be built out of our MARC records and be automatically updated when our record is altered.

There is still the concern of how much cataloging time this will take. Under the theory that some cataloging is better than none, we did a one day, one person pilot study of about 12% of the resources in our Gateway that were not already in our local catalog. Of these 100 titles, 60% of them had copy on OCLC, some having copy only days old. This percentage is not as high as I had hoped, but far higher than we anticipated. Some of the pages with no copy are of such major web sites that I am sure copy will be

added to OCLC soon. Even if no copy is located, full cataloging is easily justifiable for these important sites.

In this pilot project, we made few updates to the records. In most cases, we simply made sure the URL was correct and made sure there was access for the title as it currently appeared on the page. We added a local subject heading (IaUweb) so that we could find them and used a function key to add standard information into the location and call number fields.

We have not determined what we will do with the remaining 40% of pages with no OCLC copy. Some of these pages may be deemed ephemera, similar to vertical file materials. It may be that we will continue to have a limited database, preferably using Dublin Core, for only these materials. However, most of the selected web pages are obviously worth cataloging to some degree. As catalogers, we need to remind ourselves that we do not need to create perfect records all the time. We need to have records that are adequate for our patrons' needs. As long as we code them as non-standard records, there should not be a problem if they are shared in national databases. Ideally, there will be a consensus among libraries about what fields are critical to patron access. This will allow better sharing of these non-standard records.

While our solution may not be ideal, we believe it is workable for now. We believe that we can provide better access to selected web pages than we currently have by providing less than perfect cataloging. We feel it will be easier to train catalogers to make minimal records in MARC than to train them in a new system such as Dublin Core or to use CORC. We look forward to the time when harvesting technology has improved sufficiently to actually save us time. Dublin Core that exists in pages will be cross-walked with the OPAC so that our local web pages will be searched at the same time.

We must now determine which fields are critical for local users and make what we affectionately call "scrappy little records". The next year will be a busy one, but we hope to have records in our OPAC for all the web resources by the time we migrate next August. We have decided to spend our time working with the shared files that libraries already have (OCLC), in a format we already use (MARC), using a

simplified version of content standards we already know (AACR2), rather than expend our limited resources developing a possibly redundant proprietary database.

[1] Auburn University. (9/27/99) "Auburn Procedures for the CORC Project" URL: http://www.lib.auburn.edu/catalog/docs/corcprocedures.html [10/12/99]
University of Minnesota. (5/25/99) Content Standard for CORC Dublin Core Records" URL: http://www.lib.umn.edu/ts/drafts/Cat/CORC/contstand.html [10/12/99]

[2] Tennant, Roy. (1998) "21st-century cataloging" Library Journal. April 15:p.30-1 URL http://www.bookwire.com/LJDigital/diglibs.article$8260 [10/8/99]