User-Generated Metadata for ETDs: Added Value for Libraries

Sharon Reeves

Library and Archives Canada, Theses Canada

Abstract

Theses Canada, the Canadian national theses program, was established at the National Library of Canada[1] in 1965 before computers became ubiquitous in the home and workplace and the Internet was just a dream on the horizon. Its goal was and continues to be twofold: to facilitate access to theses and dissertations approved by Canadian universities and to preserve them in the Library and Archives Canada (LAC) collection. In 2002, as the result of a consultation with Canadian universities, Theses Canada undertook to develop a national electronic theses program based on the principle of open access. In January 2004 the Theses Canada Portal was launched, providing access to over 45,000 Canadian electronic theses and dissertations (ETDs) acquired from ProQuest, which began to digitize them from print theses in 1998 under the terms of its contract with LAC. However, in order to more fully integrate ETDs into the national program on an ongoing basis, Library and Archives Canada concurrently developed the capacity to harvest both ETDs and metadata from universities using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The longevity of the national theses program provided a firm foundation for collaboration between LAC and its university partners in building the interoperability infrastructure that the OAI-PMH requires. LAC is currently harvesting metadata and ETDs from seven Canadian universities and expects to harvest other universities as they establish their ETD programs. While Theses Canada's primary objective in developing its harvesting program was to provide open access to Canadian ETDs it has had unexpected fringe benefits. Graduate students supply the initial metadata for their theses and dissertations at the time they submit the final versions to their universities. This paper will discuss the added value service that user-generated metadata for ETDs provides to libraries.

History of Theses Canada

In 1965 a centrally coordinated theses program was established at the National Library of Canada (NLC). This was before the time that computers became ubiquitous in the home and workplace and the Internet was just a dream on the horizon. The program was established at the request of various deans of Canadian graduate schools. Theses and dissertations at that time would often end up on the shelves of university libraries gathering dust and the deans of the graduate schools wanted the NLC to develop a program which would facilitate access to documents by identifying what was available and by providing copies to researchers via interlibrary loan or sale. The Canadian Theses Service made it easier to locate Canadian theses and dissertations by cataloguing them and listing them in *Canadiana*, Canada's national bibliography, and in a bibliography called *Canadian Theses*. Preservation was also a significant component of NLC's program.

For thirty seven years, until 2002, the theses program continued to deliver its mandate in the traditional way. Print theses and dissertations were reproduced on microfiche and, starting in 1998, digitized by ProQuest Information and Learning under the terms of NLC's contract with it. By the end of 2002 the theses and dissertations in the NLC collection numbered over 220,000. It now numbers over 260,000. Theses Canada continues to deliver its traditional program but in 2002 started to develop a concurrent program for electronic theses and dissertations.

[1] Now part of Library and Archives Canada

Implementation of ETD Programs in Canada and Elsewhere

In 1992 the Coalition for Networked Information (CNI) organized a meeting to explore the feasibility of a project to capture and store electronic dissertations. This led to the development fledgling programs for student submission of electronic theses at some U.S. universities, notably Virginia Polytechnic Institute and State University and the University of West Virginia. In 1996 the Networked Digital Library of Theses and Dissertations (NDLTD) came into being, with the objective of promoting the adoption, creation, use, dissemination and preservation of electronic analogues to the traditional paper-based theses and dissertations. The NLC was represented at that first meeting coordinated by CNI and has been an active participant on the NDLTD since its inception first on its Steering Committee and now on its Board of Directors.

During the late 1990's a number of Canadian universities started to develop an interest in starting up electronic theses submission programs for their graduate students. In December 2000 the NLC hosted a national consultation to determine how to create a national ETD program. Seventy-eight participants from across Canada attended, including deans of graduate schools, university administrators, library professionals and graduate students. The outcome of the consultation was that Theses Canada was given a mandate by Canadian universities to develop an electronic theses program in order to ensure that Canadian ETDs are openly accessible in Canada and around the world. Out of this was born the Theses Canada Portal.

The Theses Canada Portal

By 2002 a systems proposal for the development of the Portal had been prepared by NLC staff. At a meeting of the Theses Canada Advisory Committee in October 2002 the proposal was endorsed. In 2003 the Portal was designed, a search interface was created, content was added and the Portal was seeded with the 45,000+ theses and dissertations digitized by ProQuest up to August 31, 2002. It was launched at the Ontario Library Association conference in January 2004 and was an immediate success at Canadian and international universities. The Portal can be accessed at www.collectionscanada.ca/thesescanada.

At the same October 2002 meeting the Theses Canada Advisory Committee decided to conduct a pilot project to acquire ETDs directly from two universities based on the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) in order to more fully integrate ETDs into the national program on an ongoing basis. As Kathleen Shearer explains in her paper on the Open Archives Initiative the protocol is a set of interoperability standards that can be used by a variety of communities engaged in publishing content on the Web. Any network server can use it to create metadata to describe objects located on that server and make it available to service providers or other repositories that wish to collect the metadata (Shearer 2002). The advantages of using the OAI protocol to set up the ETD program at the NLC were clear. Universities interested in starting up ETD submission programs intended to make those programs OAI compliant so that other organizations, such as the NDLTD, could harvest their data repositories.

Concurrent with the development of the Theses Portal, over the course of 2003 staff at the NLC developed an OAI harvester and data repository. In developing its harvesting program the Library faced a challenge that most institutions do not have. In order to acquire and preserve a comprehensive collection of theses and dissertations, thus fulfilling the NLC's mandate, it was necessary to harvest not only theses metadata but the electronic theses and dissertations as well.

This was accomplished by writing an extension to the harvesting application to allow for the harvesting of the e-theses themselves.

The universities that originally participated in the pilot project were the Université Laval and the University of Waterloo, both of which were early developers of ETD submission programs for their students. Later the University of Saskatchewan and the University of Manitoba were included.

The pilot project was intended to run for a period of six months, from April 1 until September 30, 2004. In fact the project lasted a year, until March 2005. Over that time period Library and Archives Canada harvested 693 metadata records, in both Dublin Core and ETD-ms (Electronic Theses and Dissertations metadata standard) format[2], into its data repository These records were in turn harvested by the NDLTD Union Catalog, which had already harvested the 45,000+ records already acquired from ProQuest. At the end of March 2005 the harvested records were converted to the MARC 21 format (the standard metadata format used in library catalogues) and uploaded to AMICUS, LAC's online bibliographic database.

Analysis of Harvested Metadata

LAC staff had some concerns about what the converted records might look like so after they were uploaded the records were analyzed to determine their level of quality and to assess how well they conformed to international cataloguing standards. A few bugs were identified, most of which were easy to fix. The one bug that has proven impossible to eliminate is that symbols and punctuation in some abstracts do not convert from ETD-ms to MARC 21 properly. LAC programmers attempted to solve the problem in the original set of harvested records by changing the conversion specifications for each and every error but the problem proved to be too extensive. For this reason Theses Canada notes in its harvesting requirements, which are available on the Theses Canada Portal, that in the conversion to MARC 21 the loss of some special characters, formulae and math coding will be unavoidable.

What is User-Generated Metadata and How Does it Work?

So what exactly is user-generated metadata? When a student is ready to submit the final version of his or her thesis to the faculty of graduate studies he or she is required to fill out an online template with basic metadata tags such as name, thesis title, degree, degree date and abstract. Although the students do not realize it they are creating the basis of the metadata records for their theses. The student then uploads the template and his or her thesis to the Faculty of Graduate Studies (FGS) at the university. After the FGS approves the thesis it is then released to the university library along with the template containing the basic metadata tags. From this point the process is largely automated. When the student template is sent to the library it goes through an automated conversion process that turns it into Dublin Core and ETD-ms metadata records, which are located in an OAI data repository. From there the metadata is available to be harvested by union catalogues and other institutions. The records require little intervention on the part of library staff. Of course the process varies from university to university.

---

[2] The ETD-ms is based on Dublin Core but includes extra elements specific to theses, i.e., degree name, level, discipline and grantor.

Library and Archives Canada harvests university data repositories on a monthly basis. When the metadata arrives in LAC's data repository it is automatically converted to the MARC 21 format and uploaded to AMICUS and the Theses Canada Portal. From there it is accessible to LAC's clients. Because the process is entirely automated Library and Archives Canada cataloguers do not have to make any changes or adjustments to these harvested records.

ETD-ms Tags	MARC 21 Tags (tag number in parentheses)	Common Data
oai_etdms:title	Title proper (245)	User-Generated Metadata for ETDs: Added Value for Libraries
oai_etdms:creator	Author (100)	Reeves, Sharon
oai_etdms:subject	Index term-uncontrolled (653)	Library and Information Science
oai_etdms:subject	Index term- uncontrolled (653)	ETDs
oai_etdms:subject	Index term- uncontrolled (653)	User-Generated Metadata
oai_etdms:description	Abstract note (529)	Theses Canada … metadata for ETDs provides to libraries.
oai_etdms:publisher		University of Ontario
oai_etdms:date		2007
	Place of publication: Publisher, Date (260)	Ottawa: University of Ontario, 2007
oai_etdms:type	Index term-genre/form (655)	Electronic Thesis or Dissertation
oai_etdms:format	Type of computer file (516)	application/pdf
oai_etdms identifier	Electronic location & access (856)	http://www.collectionscanada.ca/obj/s4/f2/dsk3/OONL/TC-OONL-123.pdf
oai_etdms:identifier	Electronic location & access (856)	http://etd.uontario.ca/etd/sreeves2007.pdf
oai_etdms:identifier	Local call number (099)	TC-OONL-123
oai_etdms:language	Control field (008-contains other data)	en
oai_etdms:rights	Terms governing use and reproduction note (540)	Copyright: 2007, Reeves, Sharon. All rights reserved.
oai_etdms:(degree) name		Doctor of Philosophy
oai_etdms: (degree) discipline		Library and Information Science
oai_etdms: (degree) grantor		University of Ontario
	Dissertation note (502)	Thesis (Doctor of Philosophy)—University of Ontario, 2007
No ETD-ms equivalent	Bibliography note (504)	Includes bibliographical references.

Table 1: This is an example of a metadata record harvested by Library and Archives Canada. It indicates the common data elements and compares the ETD-ms and MARC tags. Text in blue represents data that the student filled in on the submission template. Many of the other tags were generated automatically. The record above is not complete. Some control fields have been left out. Please note that the data in the table has been made up.

Benefits of User-Generated Metadata

While Theses Canada's primary objective in developing its ETD harvesting program was to provide open access to Canadian electronic theses and dissertations it has had an unexpected fringe benefit. LAC has realized significant cost savings that result from not having to create a catalogue record for each and every electronic thesis and dissertation acquired. It costs Library and Archives Canada $29 to create an abbreviated level cataloguing record. This means that, by not having to make changes to the harvested records, the institution saved $95,000 in the 2006-2007 fiscal year and will save progressively greater amounts in the future as more and more universities start up ETD submission programs for students. University libraries can similarly reduce the cost of cataloguing their theses and dissertations.

While the metadata records do not entirely conform to the international descriptive cataloguing standards in the *Anglo American Cataloguing Rules, second edition*, for example words in titles include capital letters, the level of detail they contain is sufficient to make them very good quality access records. Clients who don't particularly care about the niceties of cataloguing can easily find the titles they are seeking. As a result clients anywhere in the world can locate records for titles of interest to them either on the Theses Portal or in AMICUS and can access the ETDs on LAC's server. As well, metadata for LAC's ETDs can be found in the NDLTD Union Catalogue and via the Scirus and Google Scholar search engines.

Disadvantages

As the national library of Canada Library and Archives Canada has a responsibility both to participate in the development of and to uphold international cataloguing standards. Library and Archives Canada has sidestepped the standards issue by making the records for its harvested ETDs available in its online catalogue but not distributing them in its bibliographic products, such as *Canadiana,* or through its MARC Records Distributions Service. These services make cataloguing copy available to other libraries that expect the records to adhere to the standards outlined in the *Anglo American Cataloguing Rules, second edition*. Aside from the fact that user-generated metadata records do not totally conform to international standards there are no other apparent disadvantages to using them.

Some other factors to take into consideration when setting up an ETD submission program for graduate students, although not strictly speaking related to metadata, are the up-front time expenditure required to set up an OAI harvester and data repository, particularly for IT staff, and the need for adequate storage space on a server. However the cost of these initial requirements are miniscule in comparison to the potential savings to be gained by eliminating the need for technical services staff to catalogue the harvested electronic theses and dissertations.

Current Status of Metadata Harvesting at LAC

LAC is currently harvesting metadata and electronic theses and dissertations from the following universities:

- University of British Columbia

- Université Laval
- University of Manitoba
- University of New Brunswick
- University of Saskatchewan
- University of Victoria
- University of Waterloo

It is currently testing data from Queen's University and should be in a position soon to harvest its metadata and ETDs.

A number of universities are actively moving forward on electronic theses submission programs for their students and should be ready to be harvested by LAC within the 2007-2008 fiscal year. These include the Atlantic Veterinary College (University of Prince Edward Island), McGill University, Memorial University, the Université de Montréal, Mount Saint Vincent University, the University of Ottawa and the University of Toronto. Some of these universities are of a significant size and LAC can anticipate harvesting large numbers of ETDs as they begin to participate in the ETD program.

Generally speaking when a university starts up an ETD submission program for its students the program is voluntary for the first several years and then becomes mandatory. To date in Canada the Université Laval, the University of Waterloo and the University of Saskatchewan are the only universities that require electronic submission of theses and dissertations from all graduate students. This has an impact on the number of ETDs that LAC is able to acquire through harvesting each year. It can be expected that this number will increase in several ways:

1. as universities implement mandatory submission for their graduate students;
2. as LAC is able to harvest additional universities; and,
3. as universities make available theses that they did not submit to the traditional program.

In the past some universities did not submit masters theses to Theses Canada due to the cost of the publishing fee charged by ProQuest Information and Learning. Since there is no cost to participate in LAC's harvesting program those universities have been making electronic masters theses available on their servers for LAC and various union catalogues to harvest.

More harvested metadata for electronic theses and dissertations translates into greater savings for Library and Archives Canada by totally eliminating the cost of cataloguing them.

Conclusion

This paper has demonstrated that a significant value added benefit of ETDs is the money libraries can save by not having to catalogue them. This is accomplished by using the metadata that originates from student submission templates. The metadata then undergoes one or several automated conversions in which the appropriate data elements are added to the records to convert them to Dublin Core, ETD-ms or MARC format records. This process results in the availability in library online catalogues of records that require little or no intervention on the part of the library's technical services staff. The salary savings ultimately far outweigh the initial costs of setting up an ETD submission program for graduate students.

References

Shearer, K. (2002). The Open Archives Initiative: Developing and Interoperability Framework for Scholarly Publishing. CARL/ABRC Backgrounder Series #5. Retrieved April 25, 2007, from http://www.carl-abrc.ca/projects/scholarly_communication/scholarly_communication-e.html.