



CIDL NEWS

CIDL News #9, September 2003
ISSN 1488-2000



On the Agenda



November 2002 Open Meeting sparks Godolphin Report

The viewpoints expressed at the November 2002 CIDL Open Meeting set off an organizational review. The Steering Committee decided at its January 20, 2003 meeting (Simon Fraser University) to hire a consultant to re-assess and re-evaluate the directions set for CIDL. It was seen that the environment has changed considerably since the organization was founded in 1997. The CIDL Web-site provides the Steering Committee minutes.

On March 10, the Steering Committee held a special meeting (Ottawa) and hired Jocelyn Godolphin & Associates to prepare a report on possible organizational options available to CIDL. The final report was submitted to the Steering Committee at the end of April, 2003.

The CIDL Steering Committee will discuss in detail the report's suggested options at its September 18, 2003 meeting slated for Ottawa. The report is posted to the CIDL Web-site. <http://www.nlc-bnc.ca/cidl/newe.html>

NEW CIDL SPONSORSHIP CATEGORY

The CIDL 2003-2004 membership year marks the offer of a third type of participation, Sponsorship. For an annual fee of \$1,500, sponsors receive promotional advantages and access to the CIDL ListServ. No voting privileges are available. Encourage your digital connectors to take part. Contact CIDL at: cidl-icbn@nlc-bnc.ca.

OUR ROOTS / NOS RACINES

Canada's Local Histories Online Project brought about 4,000 full-text titles on-line in Year Two. Check out the site for the new K-12 education kits.

Of the 50 participants at the CIDL Open Meeting hosted by Chair Claude Bonnelly in November 2002, four were from the Our Roots team. Seen here from the Université Laval and University of Calgary are (l-r): Université Laval: Claude Bonnelly, Library Director and Guy Teasdale, Librarian & Electronic Documents Advisor. University of Calgary: Jackie Bell, Our Roots Project Manager and Frits Pannekoek, Director of Information Resources.

www.ourroots.ca



IN THIS ISSUE:

Digitization Standards from four projects:

Linda Pearce, University of Calgary

"Metadata for Canada's Local Histories"

William Wueppelmann, CIHM

"Metadata & Early Canadiana Online"

Rida Benjelloun, Pierre Lasou, Université Laval

**"Metadata in the Université Laval
Theses & Dissertations Project"**

**"Metadata & Resource Discovery,
Government of Canada
Role of Library & Archives Canada"**

Deane Zeeman, Katherine Miller-Gatenby, L&AC

Metadata Forum, September 19 & 20, Ottawa

Profile @ CIDL:

Tim Au Yeung, University of Calgary

Inventory of Canadian Digital Initiatives News

OUR ROOTS / NOS RACINES www.ourroots.ca METADATA FOR CANADA'S LOCAL HISTORIES

Linda Pearce Manager, Library Systems,
Info.Technology Services University of Calgary

'Our Roots / Nos racines' is an exciting Canadian digitization project undertaken by University of Calgary Press and l'Université Laval Library. Its purpose is to make available online the most complete collection possible of French- and English-Canadian local histories. Uses for these often-forgotten literary gems are multitude; the fields of history, agriculture, architecture, genealogy, religious studies, and economics are some that will find here a rich source of information and inspiration.

To make the local histories available over the web has engaged the project's National Editorial Board and staff in a number of discreet activities: fund-raising, prioritize for digitization, coordinate participating institutions, software development, hardware acquisitions and support, staff hiring and training, acquisition of materials, copyright clearance, scanning, archiving, web design, uploading, and the provision of metadata.

Without the last of these – critical metadata – none of the materials would be searchable or usable. Metadata provides the hooks with which we can fish out what we need from any collection of digitized objects, whether the objects themselves are books, slides, airphotos, videoclips, or other formats.

The project metadata provides the mechanism to manage hundreds of thousands of individual page images and to manage the workflow of the entire project. Metadata is usually divided into three areas: descriptive, administrative, and structural.

The **descriptive metadata** for the *Our Roots / Nos racines* project is based on the Dublin Core standard, AACR2 II and the IMS standard, along with local needs for special tagging. In the definition of the metadata standards and cataloguing practices, there is close cooperation between the Digitization Manager, other staff in the University of Calgary's Information Resources' Information Technology Services group, and cataloguing staff. Where possible, descriptive metadata is drawn from library catalogues such as the Library & Archives of Canada catalogue, the Université Laval catalogue, and the University of Calgary catalogue. Items for which records cannot be found have their metadata created and entered by trained cataloguers.

Pearce, cont'd. Page 8.

Profile @ CIDL Tim Au Yeung

**University of Calgary Information Resources,
Manager, Digital Object Repository Technology**

Tim Au Yeung, as the manager of digital object repository technology at the University of Calgary, has a central role in the operations of the Digitization Centre at the University of Calgary. The centre successfully launched the Alberta Heritage Digitization Project (www.ourfutureourpast.ca) and the Canadian Local Histories Online site (www.ourroots.ca) in conjunction with partner institutions across Canada. Tim directs the development of the technical infrastructure needed to support these massive digitization projects (Our Future, Our Past currently has almost 1 million images online and Our Roots has 3000 volumes online consisting of 625,000 pages of digitized text) as well as the development of the technical standards and protocols involved.

Prior to joining the University of Calgary, Tim Au Yeung was the supervisor of Product Mastering and Quality Assurance for Visual Content Development at Adobe Systems Incorporated and EyeWire, Incorporated. During his tenure at Adobe Systems Incorporated, he developed the quality assurance team and protocols for Visual Content Development from the ground up.

Many learning institutions successfully partnered with the **Our Roots / Nos racines** project. The list is available at: www.ourroots.ca and www.nosracines.ca.

Project references for metadata information:

Metadata standards for digitization projects
<http://www.ucalgary.ca/~pearce/metadatastandards.htm>

Guidelines for creation of metadata

<http://www.ucalgary.ca/~pearce/metadataguidelines.html>

Technical standards for the Our Roots project (in English)

<http://www.ourroots.ca/e/tech1.asp>

Metadata and Early Canadiana Online

William Wueppelmann, Electronic Systems Specialist
Canadian Institute for Historical Microreproductions



The Canadian Institute for Historical Microreproductions (CIHM) recently changed the way it views and handles metadata for Early Canadiana Online (ECO). ECO is located on the Web at <http://www.canadiana.org/>.

Our new approach includes new goals, standards, and methods for handling metadata.

Our previous metadata format was SGML-based but, due to lack of proper validation, inconsistently-applied standards, and a variety of practices and methods for generating and editing records, it became what amounted to a special-purpose format, reliably readable only by our database application, and even then not always so reliably. Our records were also sparse, containing little information other than what was immediately useful for the original application.

In designing a new format and practices, we tried to address these issues. Our new metadata records are XML applications and are validated against an in-house developed Document Type Definition (DTD). In developing our format and practices, we addressed a number of issues: purpose, content, format, maintenance, and preservation.

We designed our records first and foremost as data repositories. The ability to cleanly and accurately store information is more important than the suitability of the format for any particular application. XML documents can be easily converted, abridged, and transformed to create other documents that are suitable for use with a particular application. Our new records are not directly

compatible with our existing database application, but it is a simple step to derive production records that our database can index and manipulate. We expect to switch to a new application in the near future. This application also will not use the native metadata records as-is, but it will be easy to use them to generate records which it can use more efficiently.

On the issue of content, our general philosophy is,
“when in doubt, leave it in.”

XML records consume little space relative to the ECO project as a whole. Our entire metadata library consumes about 4 GB of disk space, compared to over 120 GB of image data. It is easy to go back and remove unwanted information from records. It is much harder to go back and add in new information that one suddenly has a use for. We therefore include in our metadata records everything which might one day be useful and which can be obtained with a reasonable amount of effort. We also take care to preserve granularity; it is easy to amalgamate fields together but difficult to take them apart, so we would prefer to keep the information in pieces and assemble it when needed rather than put it all together only to find that we later have a need for part of what has become a single field.

Among the things we include in each metadata record are the full MARC record for the item (tagged according to field and subfield), the full OCR text, tagged according to the page images from which each block of text was obtained and according to which language settings were used to acquire that text, and information about each image, including resolution, format, and dimensions in pixels. Much of this information, such as many of the control fields of the MARC record, are of no immediate use, but it is both safer and easier to include them now than it is to pick and choose the fields we want and hope that we chose well.

Our metadata records are stored as XML documents which are designed to validate against our DTD. This makes it easy to ensure that each record is complete and structurally sound when it is generated, and after each modification. XML is an open format that is supported by a large and growing number of applications and programming and scripting interfaces (APIs). This makes our data highly portable across computer architectures, operating systems, and applications. This means that we should easily be able to use it for new purposes and in new contexts, should the need arise.

Another reason we chose valid XML for our records is that it is easy to maintain. Our records are generated and modified by a number of small Perl programs we created, but they could also be maintained using other custom, commercial, or open-source tools. After each document is created or modified, we validate it against our DTD to ensure that it is correct.


Wueppelmann, cont'd. Page 4.

Wuepplemann
From Page 3

While this process cannot eliminate the possibility of incorrect data produced by human error, it does prevent syntax errors, missing or out-of-order fields, and other structural errors from creeping in to our metadata.

Should we decide, at some point, to retire our current format, we are confident that we can perform an error-free automated migration to whatever new format we select or design.

We do not view our metadata as a static entity. We expect the content to change over time. For example, as OCR technology improves, we may decide to re-scan many of our images in order to get better quality text recognition. Cataloguing errors will be found and need to be corrected, and other changes will need to be made. Our goal is to ensure that our metadata remains available in its most up-to-date form.

We have a central repository where we store authoritative copies of each record. All changes to records are made in this repository, and the changes are copied to other copies of our records as needed. In addition to the main repository, we maintain an up-to-date copy of our metadata on the ECO server, located in a different part of town. We also make periodic backups to CD-ROM so that we will have a reasonably up-to-date portable copy of the data. 



Inventory of Canadian Digital Initiatives

<http://www.nlc-bnc.ca/initiatives>

Digital Projects recently submitted by CIDL members

Arctic Blue Books Online

University of Manitoba, Archives & Special Collections

<http://www.umanitoba.ca/libraries/units/archives/arcticbb/index.shtml>
Unique index to 19th century British Parliamentary Papers concerned with the Canadian Arctic.

"At first, the project group worked from a small, leaky office in the basement of University College. With additional support, they relocated to larger and drier accommodation in the ladies' basement powder room. Thus was born "Lady Jane's Loo and Lab."...

Leonard Frank Photograph Collection

Vancouver Public Library

<http://collections.ic.gc.ca/vpl/>
Frank's collection is the first large collection acquired by Vancouver Public Library and provides the foundation upon which the Library's Historic Photograph Collection is built.. Frank was active as a photographer in the years 1910-1944.

Hospital Architecture in Montréal

McGill University Digital Collections Program

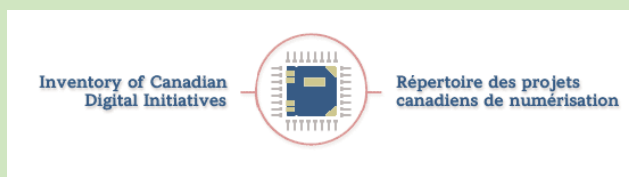
<http://digital.library.mcgill.ca/hospitals/>
Spanning 358 years, from 1642 to the present, this site contains over 1000 images, a hospital construction chronology, architectural records and an in-depth study of the Royal Victoria Hospital.

Urban Planning

McGill University Digital Collections Program

<http://digital.library.mcgill.ca/urbanplanning/>
The Urban Plan Collection is a database of reports and plans for urban and rural areas from all provinces of Canada dating from the late '50's to the present time. The value of this collection, built-up through the years as a pedagogic tool by the McGill University School of Urban Planning, is that it serves as an historic record of the shaping of urban and rural landscapes in Canada and in some cities of the United States.

This inventory provides information about and connections to projects that provide Web-accessible resources. All subjects are covered. Entries from public, private and non-profit organizations as well as from individuals are all welcome. Entries for Web-accessible resources created outside Canada that deal with Canadian topics are also encouraged.



Metadata in the Laval University Theses and Dissertations Project

Rida Benjelloun, Digital Information Management Consultant
Pierre Lasou, Electronic Documentary Resources Specialist
Université Laval Library



The Université Laval collection of theses and dissertations was launched in November 2002 (<http://www.theses.ulaval.ca>). The archive format of these theses and dissertations is XML (eXtensible Markup Language). The dissemination formats are XHTML and PDF.

The technological choices in terms of metadata for the Université Laval theses project were made based on international standards. One of the well-known metadata standards being used at present is the Dublin Core (NISO Z39.85 and ISO 15836). It provides for 15 metadata elements that can be expressed in various formats such as HTML, XML, or RDF (Resource Description Framework).

For Laval University's theses and dissertations, we use a standard derived from Dublin Core: ETD-ms.

Metadata Formats

EDT-ms (Interoperability Metadata Standard for Electronic Theses and Dissertations [version 1.00 <http://www.ndltd.org/standards/metadata/current.html>]) is an international standard for thesis and dissertation metadata, developed by NDLTD (Networked Digital Library of Theses and Dissertations).

This standard is in the form of a simple XML scheme. It is based on certain elements of Dublin Core and it provides for other elements specific to theses and dissertations. Essentially, these concern the degree: name, level, discipline.

In the absence of international standards for thesis and dissertation document models, we decided on a temporary solution and chose to use DTD (Document Type Definition) XML DocBook. (www.oasis-open.org/committees/docbook/).

ETD-ms is the metadata cornerstone of the Université Laval Theses and Dissertations Project. We can derive other metadata formats from XML metadata records:

- Dublin Core for the XHTML versions of theses and dissertations,
- Metadata Elements included in XML versions.

Indexation, Research and Dissemination

The ETD-ms metadata files are indexed by the Inktomi search engine that allows indexation of each XML element. Research in the theses and dissertations collection is considerably enhanced by this method. In fact, it is possible to perform searches using pre-defined fields (author, title, French summary, research direction, proposed degree, etc.).

To give this collection higher visibility, we became involved in the Open Archives Initiatives ... which developed a protocol ...

To give this collection higher visibility, we became involved in the Open Archives Initiatives (<http://www.openarchives.org>), which developed a protocol for exchange and dissemination of metadata. The protocol provides for two categories of participants: the "data provider" and the "service provider".

For the electronic theses and dissertations, the Library participates as a "data provider". The chosen application for implementation of the OAI protocol is the one developed by OCLC: OAICat (<http://www.oclc.org/research/software/oai/cat.shtm>).

Within the framework of the OAI, several "service providers" harvest the collection:

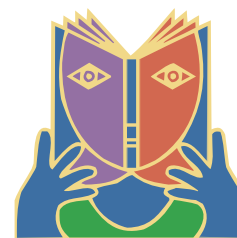
- "XTCat Experimental Thesis Catalog" of the OCLC: <http://alcme.oclc.org/ndltd/index.html>
- Public Knowledge Project's metadata archive <http://www.pkp.ubc.ca/harvester/>

Metadata Management

Metadata can require updates or corrections. It is important, therefore, to anticipate a system that will allow for this eventuality. At present, the XML records are stored in a tree directory. We are attempting to move toward managing the metadata in an XML database. 📖

Metadata and Resource Discovery in the Government of Canada

Role of the Library & Archives Canada



Deane Zeeman, Metadata Coordinator, Bibliographic Access
Katherine Miller-Gatenby, Director, Government On-Line Task Force
Library & Archives Canada

Government On-Line (GOL)

As with many other governments (e.g. UK, New Zealand, Singapore, etc.), the Canadian government recognizes that informed citizens are key to the social and economic well-being of the country. Vital to this is the provision of easy, reliable access to information, programs and services of the federal government. Canada's approach is a cross-jurisdiction effort called the *Connecting Canadians Agenda* which includes the Government On-Line initiative (GOL). The intention of GOL is to develop the World Wide Web as a service delivery channel for the Government of Canada.

Challenge of Standards

There are over 200 departments and agencies in the federal government that have extensive Web sites to cover a broad range of subjects. To provide a coherent Government of Canada presence that includes all these sites is a challenge, but it is necessary. Early in the development of the GOL initiative, the federal government recognized that citizens need consistent, reliable access to information on the web. To this end, a standard, the "Common Look and Feel Standard for the Internet" (CL&F Standard)

http://www.cio-dpi.gc.ca/clf-upe/index_e.asp was developed and implemented.

The CL&F Standard addresses several concerns of electronic access to government information including the use of metadata to improve citizens' ability to find the government information they need on the Web. In October 2001, Treasury Board ministers approved the Dublin Core Metadata Standard (DC) as a government Information Management Standard and the CL&F adopted five DC metadata elements as core elements.

Metadata Elements

The five DC metadata elements mandated for use by the Government of Canada standard are: Title, Creator, Language, Date and Subject. The standard prescribes pre-determined sets of values to be used as content for four of the five elements. (**Title** is the actual title of the Web information resource and therefore not open to standardization.) This approach facilitates finding similar information and encourages interoperability across the whole of government. The prescribed sets of values are:

Creator: Treasury Board of Canada. *Titles of Federal Organizations* (March 12, 1998)
http://www.tbs-sct.gc.ca/Pubs_pol/sipubs/TB_FIP/titlesoffedorg1_e.asp

Language: ISO639-2 Codes for the representation of names of languages: Part 2: Alpha-3 code
<http://lcweb.loc.gov/standards/iso639-2/langhome.html>

Date: ISO 8601: *Data elements and interchange formats - Information interchange - Representation of dates and times* <http://www.w3.org/TR/NOTE-datetime>.

Canadian Metadata Forum Library & Archives Canada September 19 & 20, 2003

The Forum purpose is to bring together participants from the Canadian metadata practitioner communities, both government and non-government, to discuss common concerns and develop a common approach to using metadata for improved information discovery. The goal is to cover a wide spectrum of metadata applications.

For information on speakers and the program schedule at the **Metadata Forum**:

<http://www.nlc-bnc.ca/metaforum/>

Subject: The approach to applying controlled vocabulary to the Subject metadata element is complex and multi-layered. The first layer, a broad, high-level thesaurus known as the *Government of Canada Core Subject Thesaurus (CST)* http://en.thesaurus.gc.ca/these/thes_e.html, is prescribed as the default thesaurus for federal organizations by Treasury Board Information and Technology Standard (TBITS 39.2) http://www.cio-dpi.gc.ca/its-nit/standards/tbits39/crit392_e.asp. Additionally, many departments developed their own subject thesauri, or use internationally developed thesauri within their particular subject area. These tools provide subject coverage layers that are narrower and deeper than the CST.

Controlled Vocabularies for Information Management

Treasury Board Secretariat (TBS) recognizes that these terminology sets are useful for both resource discovery and for managing the site contents, but that their use has to be controlled. Library and Archives Canada is mandated to register all thesauri, and other controlled vocabularies used in the federal government, and to make available lists of the terminology sets available for use in the Government of Canada Web space. Ultimately, this registry will become a powerful tool, as it will relate terminology used in queries at search engines to the context of a specific controlled vocabulary.

Metadata Toolset

The tools that exist are a government-wide metadata framework, standards and pre-determined terminology sets to apply values to four of five mandatory elements; guidance to identify the fifth, a standard high-level thesaurus, and a registry of controlled vocabularies. A further tool for public access to Canadian government information is the *cluster* model. Government departments are working together to cluster their information around themes and across organizational divisions. These clusters cut across department boundaries, and often include contributions from a large number of departments and other information providers (e.g. provinces, SOAs, etc.). Clusters, through the use of three gateways, are presently used on the Government of Canada Web site at www.canada.gc.ca.

Questions Remain

However, the questions remain: how do users search the federal government Web space, and can the tools provided be made more useful? Can searching this space be made easier, more intuitive and more reliable? Certainly, the structure is there. It is robust and flexible, but research shows that, although searching is a popular resource discovery option, people feel more comfortable browsing where they can see the context as they move through layers. And, more importantly, they want to use their own terminology, not an externally imposed expert vocabulary.

To address this, the L&AC and TBS are working with other government partners to develop taxonomies to support searching from the client's point of view. We clearly need to create an over-arching terminology that will link our organizational information universe, and our cluster subjects to audiences and our users' conceptual models. This will allow navigation across and into cluster content and ensure that we identify all the content appropriate to a particular concept. In this way, users will experience searching and navigation in a consistent way. Current tests of conceptual models are proving positive, but further work needs to be done.

Through the effective use of metadata and its related standards, the Government of Canada will ensure that citizens can use the World Wide Web to find the government information they need when they need it. 

**CIDL: a sponsor of the
CANADIAN METADATA FORUM
September 19 & 20, 2003
Library & Archives Canada**

**CIDL Members to speak at Forum:
Alexander Eykelhof
Director, Information Technology and
Ontario Colleges Digital Library
The Bibliocentre, Centennial College**

**David McKnight
Digital Collections Librarian
McGill University**

The purpose of the Forum is to bring together participants from Canadian metadata communities, both government and non-government, such as libraries, archives, museums, industry, educational institutions and academia. Presentations are planned on all aspects of metadata.

Program information available:
<http://www.nlc-bnc.ca/metaforum/n11-201-e.html>



Pearce
From Page 2

The *Our Roots / Nos racines* site allows browsing and searching by author, title and subject. Users can click on a map of Canada and retrieve all items for a given province. In addition, all the local histories can be searched by full-text, and documents can be viewed in three different resolutions.

It is not a trivial exercise to settle on a final group of fields for any large project, because local requirements always prevent the simple adoption in whole of any of the existing schemes. Our designed metadata allows for the development of far more sophisticated search and retrieval mechanisms.


Flexibility is the most important factor in both the metadata design and the care in which metadata is created for this project. We made sure, by adhering to well-established standards, that what we create can be ported or converted easily to other databases or metadata schemas. We also allowed for the inclusion of a wide variety of formats, so that the metadata can easily be extended to handle such items as art slides. We wrote crosswalks to define the relationships between our final metadata set and the fields in Dublin Core, CanCore, IMS and MARC.

Administrative and structural metadata are locally-defined, and provide the ability to manage the digital objects along with the intensive workflow of such a large project. Additionally, the Copyright Officer at the University of Calgary uses an application called RightsFlow, built on a Remedy platform, to handle requests for copyright clearance. Since much of the metadata in this Remedy application matches that in the *Our Roots / Nos racines* database, there is considerable discussion around building interfaces between them to avoid duplicate keying. However, most of the local histories were published a long time ago, and a large proportion of them do not need copyright clearance.

There are also markup issues. We have built index pages, or tables of contents for all the books on the site. It is possible to use TEI or another XML markup scheme to have metadata in the records to automatically build such clickable contents. Due to time and labor constraints we have not done this.

However, we are writing an implementation schema for a METS (Metadata Encoding and Transmission Standard) document. After a particular volume is located through the descriptive metadata, it needs to be presented to the user. Most of the information necessary to present each volume as though it were an online book is found in the structural metadata.

Currently, the structural metadata is based on work originally developed for the *Alberta Heritage Digitization Project*. Since the formal release of the METS standard in July, 2002 we have worked to codify the implementation of a METS-based approach to structural metadata as well as to object-metadata binding. This is done through an implementation schema that provides a more rigid set of rules to create METS XML documents, specifically in the context of the *Our Roots/Nos racines* project.

Of particular interest is the ability for a METS object to carry not only the metadata but to specify the behaviors associated with the presentation of a particular object. This ability will allow *Our Roots/Nos racines* to present documents portable to more platforms and environments than is currently possible in the existing infrastructure. 

CIDL News #9 September 2003 ISSN 1488 2000

Editor: Michelle Landriault,
CIDL Coordinator

This publication is presently
issued twice a year by the
Canadian Initiative on
Digital Libraries (CIDL).

Published articles are
copyright to the individual
authors.

Send queries or
submissions to:
cidl-icbn@nlc-bnc.ca.

To join CIDL, visit our
Web-site at:
www.nlc-bnc.ca/cidl/

Canadian Initiative
on Digital Libraries
Room 215
395 Wellington Street
Ottawa Canada K1A 0N4

