Repurposing User-Generated Metadata Pathfinder: Interim Report

March 31, 2010 (changes made April 13, 2010)

Pathfinder Working Group Members
Candace Loewen
Nancy Roberts
Bill Leonard
Risë Segall
Helen Apouchtine
Myriam McCormack
Jim Clark
Mary Carroll
Martha Coady
Gary Cleveland
Christine Alexander
George Murray
Corey O'Halloran
Joanne Paterson
Sylvie Béland
Barney Shum
Jane Thacker
Susan Haigh

Table of Contents

Executive Summary

The Repurposing User-Generated Metadata Pathfinder (hereafter called the RUGM Pathfinder) has established a path for Library and Archives Canada (LAC) to modernize its cataloguing operations. For LAC to remain relevant in a fast-changing digital universe, the RUGM Pathfinder explored LAC's ability to ingest and repurpose metadata produced by publishers about their publications.

Publishers and libraries create their own metadata in different format standards. Libraries use MARC 21 to capture bibliographic information in machine-readable format; while publishers use ONIX (ONline Information eXchange) formatted metadata, which is an international standard adopted by the book trade industry. However, the main goal of both publishers and libraries is the same: resource discovery. Rather than duplicating effort, LAC can share and enrich its data more collaboratively to effectively serve Canadians.

In Canada, ONIX is used by BookNet Canada (BNC), an agency that collects information from English-language Canadian publishers in order to share it with supply chain partners. The RUGM Pathfinder tested crosswalks from ONIX to MARC 21 and the automated transfer of publisher-supplied metadata into the LAC description workflow. However, the RUGM Pathfinder is just the first phase. In terms of success, the RUGM Pathfinder has met its initial goals (listed below). In terms of understanding the possibilities for permanently changing LAC's processes for cataloguing trade publications, the RUGM Pathfinder has surpassed its goals. Thus, this report is not only a summary of findings to date, but also—and perhaps most importantly—a blueprint for future phases of this work.

The strongest recommendation at this interim stage of the RUGM Pathfinder is for LAC to continue on the path it has found.

RUGM Pathfinder Goals

The RUGM Pathfinder sought to explore four new avenues for LAC to describe its collection using metadata created by other sources, in this case, Canadian publishers:

1. Could LAC establish a legal agreement with BookNet Canada,[1] a repository of ONIX metadata from most English-language Canadian publishers, to import metadata into AMICUS?

[1] Information about BookNet Canada (BNC) is available on their website at: www.booknetcanada.ca

2. Could LAC convert the ONIX metadata into the international library standard used at LAC, MARC 21?
3. If yes, how would the workflow change for the acquisition and description of published heritage at LAC? If yes, would LAC's relationship with publishers change?
4. Could creator-produced ONIX metadata assist clients in resource discovery?

RUGM Pathfinder Main Findings

Referencing the four questions above, as well as one general systems consideration, the main findings of the RUGM Pathfinder are the following:

1. the establishment of a legal agreement with BNC is beneficial to both parties;
2. the migration and repurposing of ONIX metadata for description of published heritage is both possible and desirable;
3. this RUGM Pathfinder must become an operational process at LAC, and common practice in interactions with Canadian publishers; more analysis must occur before this can happen;
4. the possibilities for repurposing creator-produced metadata are many, notably two important ones:
    a. less basic description (and less duplication of effort) by LAC;
    b. enhanced descriptions for resource discovery; more exploration and analysis, as well as system improvements, are needed before this possibility can be fully leveraged.

RUGM Pathfinder Recommendations

Based on the findings at this interim stage of the RUGM Pathfinder, the recommendations for the way forward for LAC to proceed are:

1. To analyze the repurposed descriptive information in the AMICUS system and note what remains to be done (such as level of quality assurance required);
2. In cooperation with BNC, to develop accurate and complete information in fields relevant to library cataloguing (MARC standard);
3. To analyze the impact of repurposed data on the cataloguing workflow, and make the necessary (irreversible) workflow changes and system improvements in order to benefit from repurposing publisher metadata;
4. To analyze LAC's operational points of contact with publishers to seek and implement efficiencies (e.g., a publishers' portal);
5. To pursue publisher metadata from all Canadian publishers to assist LAC in bibliographic description; work with those publishers (such as Francophone publishers in Quebec) who do not send their data to BNC;
6. To build on the Memorandum of Understanding between LAC and BNC to establish a legal agreement for receiving ONIX metadata from BNC;
7. To move beyond hypothesis to explore possibilities and analyze ONIX metadata for resource discovery at LAC, and then make changes to system views and interfaces, as appropriate, to optimize enhanced ONIX metadata for client resource discovery;

See also Annex 1: Potential efficiencies from operationalization of the Repurposing User-Generated Metadata Pathfinder project, which could lead to more recommendations for the way forward.

Limitations

While the RUGM Pathfinder provides great hope for the future of description and discovery at LAC, it is very important to understand the limitations of the above recommendations. They include the following:

1. The online catalogue AMICUS, like other bibliographic systems of its age and design, is an impediment to realizing the potential benefits of user-generated metadata. Modern library catalogues provide greater flexibility and ease-of-use for integrating data from a variety of sources. By further developing the data, the resource discovery experience for users will be enhanced. However, the metadata will not be used efficiently until the systems that LAC uses are modernized to provide the necessary capabilities (global updates, ease-of-use in matching and merging data, etc.). Similarly, the benefits of repurposing that data will not be realized until an optimal interface for resource discovery is implemented to permit users to view the information organized in ways that engage them and enriches their knowledge of the LAC collection.
2. Many of the benefits to LAC noted in this report are predicated on serious re-engineering of its current processes, including its relationship with publishers.
3. The extent to which BNC can work with publishers to produce quality metadata will impact on the time and effort needed at LAC to filter data and perform quality assurance on them.
4. Resource challenges at LAC are currently affecting the momentum of this RUGM Pathfinder.

1. Introduction

ONIX is an international standard used by the publishing industry to represent product information in a form that can be communicated electronically[2] across the supply chain. The MARC 21 format is an international standard used by libraries to store and communicate bibliographic data in library databases and networks. The Repurposing User-Generated Metadata Pathfinder (RUGM Pathfinder) adapted and developed processes to bridge the separate "languages" of ONIX and MARC 21 within the system environment at Library and Archives Canada (LAC) and, in doing so, to explore the opportunities and benefits of repurposing publisher metadata within LAC's operations.

The RUGM Pathfinder is comprehensive and expansive, involving major exploration in significant areas of LAC and in the business lines of acquisition and resource discovery. These areas include: (1) changing internal systems and workflows; (2) increasing collaboration, including legal agreements, with external partners; and (3) generating future opportunities for enhanced resource discovery. The RUGM Pathfinder did more than study a question or a future possibility for LAC, it actually tested the possibility. The final results will have a lasting impact on the institution and on its relationships with partners, particularly those within the Canadian publishing supply chain.

The full impact of the RUGM Pathfinder is not yet known. More analysis of the data is required to ascertain the impact on LAC of ingesting the publisher-created metadata into its system, as well as the usefulness of the data for resource discovery. The RUGM Pathfinder revealed as many opportunities and challenges as it did concrete findings. Thus, the report has two main components: a report on the RUGM Pathfinder to date (up to and including Section 7); and considerations for the next steps to take (Section 8

[2] The ONIX standard is published and maintained by EDItEUR, the standards body for the global book and serials supply chains, and is available online at: http://www.editeur.org/8/ONIX/

and following).

The goal of the RUGM Pathfinder was to assess the feasibility of developing and implementing programs and procedures enabling LAC to:

a) Receive metadata describing Canadian publications including links to associated digital material created by Canadian publishers and provided by BookNet Canada (BNC) in ONIX format.

b) Convert and load the metadata into the LAC bibliographic database (AMICUS) as a basis for LAC cataloguing records.

As part of this RUGM Pathfinder a representative sample of ONIX records received from BNC were loaded into AMICUS to permit LAC to assess their usefulness to its Cataloguing in Publication and bibliographic description operations. At the time of the writing of this interim report, a second, larger batch of records has arrived at LAC for further analysis.

2. Context

2.1 Modernization

The opportunity to enter into the Repurposing User-Generated Metadata Pathfinder (RUGM Pathfinder) is framed by the modernization exercise within Library and Archives Canada (LAC). One of the main goals of modernization—to collaborate with others in order to remain relevant in a highly volatile world—is squarely the main goal of the RUGM Pathfinder.

The RUGM Pathfinder proved that LAC can partner with other organizations in the supply chain. LAC entered into an agreement with BookNet Canada (BNC) to ingest the metadata they collect from English-language Canadian publishers; and LAC explored other avenues related to this RUGM Pathfinder with other potential collaborators (meetings with BookShelf, ANEL/Démarque, and conference calls with the Library of Congress in the U.S.). LAC desires partnerships in order to remain relevant in the fluid, volatile world of metadata exchange; such as eliminating the need for people to rekey information and expropriating the possibilities of a range of supplemental metadata. The full test is not complete. The next phase will be modifications to workflows at LAC, and this will be done in the spirit of modernization, too, as the Librarian and Archivist of Canada stated in his paper, *Shaping Our Continuing Memory Collectively*: “to guarantee the feasibility and applicability of our new processes, we must try them out and implement them gradually, taking into consideration what we learn along the way.”

The RUGM Pathfinder straddles at least two business lines at LAC, acquisition and resource discovery, but its greatest impact is in the area of resource discovery. If description is the backbone of discovery, description that is rich for discovery and that can be achieved effortlessly and efficiently is desirable. But description and discovery cannot be optimized to their full potential without a favourable “discovery layer” link between the two. The resource discovery strategies for leveraging LAC descriptions to reach clients are currently being developed; the findings in this report can form building blocks for the way forward in that business line.

2.2 External World

For decades libraries have been developing and applying standards to permit the descriptions of their holdings to be shared among organizations, and across databases and networks. The development and application of the ONIX standard within publishing organizations has brought that industry to the point where the descriptions created for trade purposes can be shared and repurposed among different partners in the supply chain, from publishers to libraries, for the benefit of each partner's respective clients.

A number of projects have sought to take advantage of these commonalities in the descriptive work done by publishers and libraries. For example, the organizations responsible for Resource Description and Access (RDA) and ONIX announced a joint initiative to develop a common framework for resource categorization. RDA is a new international standard for resource description and access. It is being developed as the successor to the Anglo-American Cataloguing Rules. The RDA-ONIX framework will facilitate the transfer and use of resource description data in both the library and publishing communities.

In Britain, publishers send ONIX feeds to commercial bibliographic utilities, such as Bowker and Bibliographic Data Services Limited (BDS), who enhance and transform this data into the MARC 21 format. They then sell this data to institutional customers, including the British Library.

Another example comes from the Online Computer Library Center (OCLC) where they offer Metadata Services for Publishers.[3] OCLC ingests publishers' ONIX title metadata and enriches the data using WorldCat mining and mapping techniques, then returns this enhanced ONIX metadata back to the publishers for use in their systems. Because the data are then made available before publication, libraries may use this information in selection, acquisition and technical services workflows.

A final example comes from the Library of Congress, where they have been experimenting with ONIX data for several years. Most recently the Library of Congress has introduced an electronic cataloguing in publication program with several select publishers which make use of ONIX metadata. When a Cataloguing in Publication (CIP) request is made, the Library of Congress cataloguer searches the ONIX repository to see whether the data may have been incipiently recorded. The data are then modified and enhanced to become a complete CIP record. Originally a pilot project, the electronic cataloguing in publication program continues to grow as new publishers are added to the pool.

These examples simply show that national libraries and large-scale aggregators incorporate the re-use of ONIX data to improve their workflows and enhance resource discovery.

3. Objectives

The Repurposing User-Generated Metadata Pathfinder (RUGM Pathfinder) explored the ability of Library and Archives Canada (LAC) to ingest and repurpose metadata produced by publishers about their publications. The intent of the RUGM Pathfinder was to permit LAC to determine the feasibility and usefulness of loading publisher-supplied metadata to AMICUS on an ongoing basis for subsequent use in LAC's bibliographic operations and products. Up to the end of March 2010, this determination involved exploring the legal considerations and the technical feasibility. The impacts and efficiencies for operations as well as the effects on LAC’s services to users (individual clients, libraries and publishers) form the second phase of the RUGM Pathfinder.

The RUGM Pathfinder provided an opportunity to test crosswalks from ONIX to MARC 21 and the automated transfer of publisher-supplied metadata into the LAC workflow (to be tested in 2010–2011).

[3] See: http://www.oclc.org/us/en/news/releases/200954.htm

4. Scope of the RUGM Pathfinder

The Repurposing User-Generated Metadata Pathfinder (RUGM Pathfinder) addresses itself only to publisher-supplied standard metadata in ONIX XML format. For the purposes of the RUGM Pathfinder, Library and Archives Canada (LAC) collaborated with BookNet Canada (BNC) as a source of validated ONIX metadata (version 2.1 as implemented by BNC). While creator-produced, crowd-sourced (i.e., user-enhanced) and auto-extracted metadata from the document are also sources for metadata, they are not in the scope of this RUGM Pathfinder.

It must be noted that BNC deals primarily with the English-language publishing industry and supply chain. The RUGM Pathfinder did not deal with any equivalent organizations in Quebec, but future phases will include collaborating with French-language publishers. As a result most of the records loaded into AMICUS during the RUGM Pathfinder were for English-language publications.

Another constraint was due to the continuing development of BNC's new Biblioshare database which is still in beta and LAC is dependent upon them to supply a sample set of ONIX metadata. The data analysis and ingest were limited and LAC is not able to test subsequent updates with regard to record matching.

As the RUGM Pathfinder progressed other factors influenced its scope. It was decided that digital objects (such as URLs within notes) related to the publisher-supplied metadata would be referred to in bibliographic records but would not be stored or managed by LAC. Also, changes to the public interface to LAC library material (library search and subsequent item display) were excluded. And finally, programs and procedures for the ongoing loading of ONIX data (e.g., to schedule and run conversion and loading programs automatically) were excluded from the RUGM Pathfinder.

The timeframe for the RUGM Pathfinder was also a factor that significantly influenced its scope. The RUGM Pathfinder was dependent on BNC's schedule for producing files of records. Because BNC was developing and refining its own system during this same period, they were not able to produce more than one file of records during the RUGM Pathfinder, and BNC was unable to apply any filters to the records that were sent to LAC in that file. As a result, the initial phase of this RUGM Pathfinder did not deal with the important issues of matching original and updated versions of the same records, and filtering became an issue that LAC had to deal with after importing BNC's records rather than having BNC do an initial screening on the types of records it was exporting to LAC.

5. Methodology

The Repurposing User-Generated Metadata Pathfinder (RUGM Pathfinder) was conducted over the course of five months from October 2009 to March 2010. During this time period, Library and Archives Canada (LAC) acted to secure a Memorandum of Understanding with BookNet Canada (BNC) in order to obtain permission to repurpose their data. Concurrent with this effort, investigations into the feasibility of integrating ONIX data into the AMICUS system were underway. To this end, a small sample data set was obtained from BNC. The original sample set numbered only 63 records. After discussions with BNC, a larger file of 17,993 records was obtained for review. To these another 16,000 records were later added. Investigations into the best way of addressing this challenge led to the methodology described below. The next phase of this project will involve analyzing the impacts of these BNC records on LAC workflows.

The steps involved in the conversion and loading of ONIX records from BNC records were:

1. **ONIX to MARC XML (XSLT) conversion**
   Analysis was performed as to what the best methodology for migrating ONIX metadata to MARC 21 might be. It was decided the best course of action would be to adapt an existing tool created by the Library of Congress for converting ONIX to MARC XML. The Library of Congress's tool set was designed for an earlier version of the ONIX standard (version 1.2). The records furnished by BNC are version 2.1. The XSLT tool (XML stylesheet) was greatly enhanced to provide functionality for the records currently provided by BNC.

2. **MARC XML to MARC 21 conversion**
   Another XSLT transformation tool that had already been developed by LAC for the E-Theses Project was repurposed to perform the conversion of MARC XML records to MARC 21.

3. **Filtering**
   Based on the records provided by BNC and the ONIX specification, LAC developed filters for selecting which types of records to load. Various product types included in BiblioShare are currently being filtered out by LAC. For the most part these filters focus on variant formats of items that LAC does not collect (calendars, displays etc.).

4. **Further Processing**
   The ONIX data set carries several variant text formats: ISO 8859-1 HTML character sets, decimal and hexadecimal character values and HTML mark-up are included in the files. These codes are all removed by a separate process designed specifically to clean-up the data on throughput.

6. Legal considerations

The legal considerations touching on the participation by Library and Archives Canada (LAC) in the Repurposing User-Generated Metadata Pathfinder (RUGM Pathfinder) on the receipt of metadata from BookNet Canada (BNC) were straightforward in nature.

In assessing any risk applicable to an agreement, LAC must first ensure that in entering into it, it is acting within its legislative framework. In the case of the agreement with BNC, this activity by LAC did not raise any flags. There is no question that in fulfilling its statutorily mandated duties, LAC receives, on an ongoing basis, metadata from publishers on upcoming and current publications. Logically and legally, there appears to be no impediment to LAC also receiving this metadata, indirectly from publishers, via BNC.

Because the metadata being provided in the RUGM Pathfinder contains information over and above that which publishers would ordinarily supply to LAC as part of their legal deposit obligations, LAC, as a result of the RUGM Pathfinder, is now able to extract information from this metadata and to repurpose the data for other uses.

Since LAC is acting within its mandate in receiving the metadata, what legal risks might its participation in the RUGM Pathfinder entail? Theoretically, there are a few potential risks. As an initial example, if, conceivably, LAC were to inadvertently post material that infringed on copyright protection, there could be a risk to its reputation. The Memorandum of Understanding addresses this potential risk by LAC confirming in the agreement that it may be requested, in the event of any inaccuracy in material posted, to modify or remove metadata earlier received from BNC.

LAC does not have a direct contractual relationship with Canadian publishers regarding the metadata received for the RUGM Pathfinder. The metadata coming to LAC is filtered by BNC, after the latter receives it from Canadian publishers.

Since LAC has no independent or cost-effective means of verifying the accuracy and legal title to the content provided to it by BNC, the Memorandum of Understanding contains specific warranties that all intellectual property rights in the metadata provided have been cleared.

In addition to the warranties regarding intellectual property, BNC has specifically confirmed, in the Memorandum of Understanding that it has full legal right and entitlement to provide the metadata to LAC. Moreover, the Memorandum of Understanding further provides that BNC has both disclosed the provision of this metadata to LAC to all publishers, and has obtained their express or implied consent to it, prior to the metadata being provided to LAC.

In short, because of the nature of the RUGM Pathfinder, the legal considerations affecting it are discrete in nature and straightforward. There does not appear to be anything in this project that would impact on LAC from a legal perspective that has not been both foreseen in and answered by the terms of the Memorandum of Understanding governing it.

However, it should be noted that the ability of LAC to enter into a Memorandum of Understanding with BNC bears some connection to the nature of their arrangements with publishers. At present, BNC handles book cover artwork and related digital objects by leaving those materials on publishers' websites and linking to them. Due to the frequency of broken links to those objects, BNC may decide to collect and store those digital objects on its own servers. If that were the case, there could be additional issues relating to copyright and access that would have to be considered and addressed in another Memorandum of Understanding.

7. Information technology considerations

It was agreed at the outset that BookNet Canada (BNC) would attempt to apply filters to the records it sent to Library and Archives Canada (LAC) so that various types of products and/or records were excluded from the files of records received in the Repurposing User-Generated Metadata Pathfinder (RUGM Pathfinder). LAC prefers to receive ONIX records excluding materials not collected by LAC such as toys or calendars. The criteria list was supplied to BNC. In the end, however, BNC was unable to provide a filtered subset of records within the timeframe of the RUGM Pathfinder and LAC developed filters in-house (as mentioned above). It is assumed that BNC will implement filtering before ONIX loading becomes a production process.

The publisher metadata do not conform to a single or standard character set. This could cause some problems such as the program returning error values for these records in the future.

Removal of some HTML coding from the ONIX records during conversion to the MARC 21 format proved to be a challenge. This is an area where publishers could improve, but it is also an area where LAC can expect to encounter new challenges in the future.

The design of the AMICUS system was also a consideration during the RUGM Pathfinder. The segregation of acquisitions and bibliographic records in different views within the AMICUS database forced a decision to load the converted ONIX records into the cataloguing interface (commonly referred to as the NL view) where it would be easier to work on those records for cataloguing purposes. From the client perspective there are concerns because the publisher records will be searchable as part of LAC's collection in AMICUS. This will present difficulties for LAC's client services, if users assume that LAC owns the items described in those records or that LAC is responsible for the accuracy and quality of that data. Although a note will be added to each BNC record in AMICUS to specify that they were supplied by the publisher, that may not be enough and changes to the client interface to AMICUS should also be considered.

8. Implications for LAC operations

The next phase of the Repurposing User-Generated Metadata Pathfinder (RUGM Pathfinder) will be carried out in 2010–2011. It will focus on analyzing the impacts of loading a file of approximately 6,000 BookNet Canada (BNC) records into AMICUS on the workflows at Library and Archives Canada (LAC). However, based on an initial analysis of the test files that were converted to MARC 21 (see Annex 1) the potential efficiencies on existing workflows are evident. The assumed benefits and risks for LAC's operations and products, described below, will be examined further during the analysis phase of the RUGM Pathfinder.

8.1 Acquisitions operations:

a) For publications received on legal deposit, LAC will be able to use BNC records as the basis for its acquisitions records. Currently tombstone data are input manually into AMICUS.

b) For forthcoming publications with no CIP records, LAC will be able to search AMICUS directly for new titles. LAC will benefit as less resources are spent in searching publisher websites and other sources for information about upcoming publications. LAC will be able to use BNC records as the basis for requisition/claiming records. Currently tombstone data are input manually into AMICUS.

c) Further efficiencies for acquisitions will only be realized if BNC records include tombstone data that will allow LAC to send out claims with minimal human intervention (i.e., all records should include information on planned date of publication).

d) Records for forthcoming publications need to be clearly identifiable by using searching or reporting functionality. Currently, searching functionality in AMICUS is limited and reporting functionality is non-existent.

8.2 Cataloguing in Publication (CIP) operations:

a) CIP records derived from ONIX records will be created following recognized descriptive, classification and subject analysis standards so that these records can continue to be used by the library community. LAC will continue to enhance and produce MARC records for distribution to libraries.

b) For CIP requests received for publications with existing BNC records, LAC will spend less time rekeying or cutting and pasting tombstone data into AMICUS. BNC records in AMICUS will be used as the basis for CIP records.

c) The availability of enriched metadata such as tables of contents, summaries and author biographies will provide LAC with more information on the subject of the publication or background information about the author. This will eliminate the need in some cases to contact the publisher and/or author to request details.

d) Further efficiencies for LAC will only be realized if BNC records are accurate, without errors and follow the ONIX standard. If BNC records do not conform to a certain standard, it is questionable as to whether there is value in using them. Better quality data reduces costs, improves productivity and accelerates processes.

e) CIP requests are currently submitted by publishers to LAC using a form available on LAC's website. Further investigation is required to determine whether it would be possible to synchronize CIP submission requests with the submission of data to BNC in order to make the process more efficient for all parties.

f) Up to this point, LAC does not have a true sense of when information about forthcoming publications appear in the BNC database. If this information becomes available only after a publisher initiates a request for CIP data, then the usability of BNC data for CIP declines.

g) The ability to do Z39.50 protocol searches on the BNC database may also introduce a significant amount of savings for CIP. Pre-publication information is fluid and CIP is a time-sensitive process. Publishers add information to the ONIX record on a continual basis. Currently, LAC plans to load records into AMICUS in batch files. Updates to records would only be included in subsequent batch loads. Z39.50 is a client-server protocol for searching and retrieving information from remote computer databases. Using this protocol for CIP would enable LAC to search BNC when required and to find the most up-to-date information available. Eliminating batch loads would make for a more efficient AMICUS database (which is always more cost-efficient) and would require less LAC intervention in the data. The Z39.50 protocol is a common feature in most Integrated Library Systems; however, its functionality is not available in the AMICUS system.

8.3 Derived cataloguing processes:

a) MARC records derived from ONIX records will be created following recognized descriptive, classification and subject analysis standards so that these records can continue to be used by the library community. LAC will continue to produce enhanced MARC records for distribution to libraries.

b) Most description done by LAC is for new publications. Generally speaking, cataloguing copy does not exist for more recent titles. Currently LAC must manually input or cut and paste tombstone data into AMICUS.

c) For new titles received, LAC will be able to use BNC records as the basis for acquisition/cataloguing records.

d) The availability of enriched metadata such as tables of contents, summaries and author biographies will provide LAC with more information on the subject matter of the publication leading to better classification, subject analysis and authority creation.

e) Further efficiencies for LAC will only be realized if BNC records are accurate, without errors and follow the ONIX standard. If these records do not conform to a certain standard, it is questionable as to whether there is value in using them. Better quality data reduces costs, improves productivity and accelerates processes.

8.4 New Books Service:

a) New Books Service is a GenAMICUS application. It takes data from AMICUS and generates content on current and forthcoming Canadian publications for inclusion in the New Books Service website. Minimal LAC resources are required to produce this site.

b) New Books Service requests are currently submitted by publishers to LAC using a form available on LAC's website. Publishers can attach files containing enriched metadata such as tables of contents, introductions, book cover artwork, etc. In order to include these files on the New Books Service, they must be saved to LAC servers.

c) Upon further analysis, it may be determined that LAC should no longer store these files on LAC servers. This content would be available to users by linking to the publisher's website directly. As a result, there would be no further requirement for LAC resources in this process.

d) New Books Service is a collection development tool widely used by Canadian libraries. The current website; however, does not provide users with an optimal resource discovery experience and does not take full advantage of the enriched metadata provided by publishers. LAC should consider approaching BNC or Canadian Bookshelf about possibly collaborating on a joint product similar to New Books Service that would serve both the publishing and library communities. Canadian Bookshelf is an initiative of the Association of Canadian Publishers that plans to aggregate information about Canadian-authored books and provide access to these titles via a community-based discovery platform. The Canadian BookShelf platform will be populated with data received from BiblioShare and the Association is currently testing LAC bibliographic and authority records for use in their discovery platform.

9. Implications for resource discovery

During the Repurposing User-Generated Metadata Pathfinder (RUGM Pathfinder) and in the subsequent analysis that will follow, the focus has been on determining the immediate effects of publisher-supplied metadata on the operational processes at Library and Archives Canada (LAC). However, the "down the road" opportunities for enhanced resource discovery presented by the enriched publisher metadata are already clear. Once the RUGM Pathfinder is made operational, LAC will have an opportunity to use this additional metadata to enhance resource discovery in modernizing its online catalogue and aligning it with client expectations. The enriched metadata from publishers (book cover artwork, table of contents, summaries, author biographies) will provide LAC with the information required to meet the growing expectations of its clients.

As LAC moves forward with the replacement of the AMICUS system and as it develops a new online catalogue for use by its clients, LAC will need to determine the optimal manner to display the enriched metadata. Recent OCLC research demonstrates that end users of their WorldCat catalogue expect and value enhanced content including summaries/abstracts and tables of contents presented in a clear and intuitive manner:

> "Discovery-related information elements beyond author and title, such as summaries, excerpts and tables of contents, are essential aspects connecting the stages of an end user's discovery-to-delivery experience.
>
> While conducting their searches, end users value and expect evaluative information to assist their discovery, and ultimately, the delivery of materials. Libraries need to make it easier for end users to quickly ascertain whether items meet their needs; for analogue items, the available data needs to help users decide if it is worth their time to obtain the items—most often by going to the library."[4]

Users of library catalogues are increasingly expecting an "Amazon"-like experience within library catalogues. And increasingly, academic, public and national libraries are responding to this challenge.

[4] *Online catalogs: what users and librarians want: an OCLC report / principal contributors*, Karen Calhoun … [et al.] -- Dublin, Ohio: OCLC Online Computer Library Center, c2009. -- vi, 58 p. -- ISBN 1556534116. Available online at: http://www.oclc.org/reports/onlinecatalogs/fullreport.pdf

The successful implementation of this RUGM Pathfinder will provide LAC with the opportunity to move forward on this front.

10. Next steps

As already noted in the Memorandum of Understanding this work has been done as a pilot project and as such is by its nature incomplete. The converted file of BookNet Canada (BNC) records that was loaded into AMICUS must be reviewed and the impact of those records for acquisition, description and resource discovery at Library and Archives Canada (LAC) must be analyzed. An initial list of issues to be addressed in the analysis phase as well as issues for long-term consideration is given below.

10.1 Issues for the analysis phase

a) to test the assumed benefits and risks outlined in Section 8 in relation to LAC's operations and workflows for acquisition and cataloguing of trade publications;

b) to determine whether the filters applied to the types of records received from BNC were appropriate or should they be further refined to accept or reject additional categories of publications (e.g., foreign publications about Canada or by Canadian authors);

c) to determine the optimum method for acquiring records from BNC (e.g., by batch downloads or by selective harvesting);

d) to specify the desired frequency for receiving new and updated records from BNC's database;

e) to determine the volume of incoming records that would be expected in a production environment;

f) to specify the procedures that would be required to support ongoing loading of ONIX data (e.g., to schedule and run conversion and loading programs automatically);

g) to specify the automated processes that would be required to handle the receipt of updated and modified ONIX records from BNC for example, replace the original BNC record if not already modified with LAC data; reject the updated BNC record if original version already modified by LAC; load records to a location other than the cataloguing interface (commonly referred to as the NL view) of AMICUS and use simple replacement;

h) to specify the mechanisms for identifying and dealing with duplicate ONIX records in different batches of records from BNC;

i) to determine whether it is better to load publisher metadata into the AMICUS Acquisitions view (instead of the NL view) or to load them into a new location view, to be determined;

j) to specify how the LAC business areas (acquisitions and resource description) will be notified of new loads of BNC records, and which people should be informed;

k) to determine whether LAC should store and manage copies of digital objects (e.g., book cover artwork) related to the ONIX records received. If so, the legal and technological details would have to worked out between LAC and BNC (and between BNC and publishers);

l) to recommend the mechanism for providing BNC with feedback on issues of record quality, as requested by BNC.

10.2 Issues for project operationalization

Costs and associated funding to operationalize the intake and integration of ONIX metadata to AMICUS were not addressed within the scope of the RUGM Pathfinder. For the purpose of the RUGM Pathfinder, LAC acquired a single batch file from BNC and loaded it into AMICUS. To operationalize the ingest of ONIX data will require further work from LAC to develop notifications, matching algorithms and other mechanisms for handling updated and modified records. Requirements, specifications and associated costs for this work will have to be developed. The effects of acquiring duplicate and updated records in the existing systems environment and the impacts of BNC's records on LAC's workflows and products are still unknown. The analysis phase may present additional questions that will need to be addressed before the RUGM Pathfinder can be successfully integrated into operations.

Some of the information technology issues to be operationally addressed before the loading of BNC records into AMICUS are listed below.

a) the optimum method for acquiring records from BNC's database (e.g., via batch downloads or selective harvesting);

b) the desired frequency for receiving files of records from BNC;

c) developing software programs and procedures for the ongoing loading of records from BNC (e.g., to schedule and run conversion and loading programs automatically);

d) refining the filters that are applied to determine the types of records and product data received from BNC;

e) developing methods and programming to deal with updates to ONIX records that LAC has already received and converted to MARC 21 and which LAC may have subsequently modified with library data such as subject headings, Dewey classification, etc.;

f) creating a mechanism for notifying LAC when new BNC records are available (and determining which people should be informed).

10.3 Issues for long-term consideration

a) As the scope of publisher metadata is expanded to include French-language publishers in Canada (or others groups not within the scope of the current RUGM Pathfinder) further development work will be required.

b) The need to ensure long-term access to and preservation of digital objects associated with enhanced publisher metadata, such as digital images of the book cover artwork for a publication. If these digital objects are linked from the bibliographic records, then LAC should take the necessary steps to ensure that its users have stable, long-term access to these objects, instead of encountering links that have changed or no longer exist to publishers' websites.

c) The impact of a modern system replacement for AMICUS on the methods LAC employs to obtain publisher metadata (e.g., using the Z39.50 protocol to search and download remotely from BNC's database on an as-needed basis for cataloguing purposes versus batch loading of separate files of BNC records into the cataloguing module of AMICUS).

d) When and how LAC manages a transition to ONIX 3 (and future versions). Should LAC expect to continue receiving data in multiple ONIX formats? Does LAC need to be more involved in ONIX format development?

Annex 1: Potential efficiencies from operationalization of the Repurposing User-Generated Metadata Pathfinder project

EFFICIENCY	ANTICIPATED IMPACT
Less or no rekeying of basic bibliographic data by staff in Acquisitions and CIP	Less inputting by LAC staff
Streamlined workflow – one step closer to single window for publishers	Less duplication of inputting for publishers; example: New Books Service and CIP workflows are addressed
ONIX metadata available before publication or LAC receipt of item	PH Acquisitions spends less time searching for publication data
Management of digital objects is streamlined; LAC will not have to handle each object one by one	Less cutting and pasting (as happens currently) of digital information for New Books Service
Receipt of metadata about items we would not otherwise know about / acquire	Better coverage of Canadian publications; more quality assurance of bibliographic data due to more comprehensive coverage
Availability of tables of contents, bios, abstracts, summaries will enable subject analysis early on in description	Information to perform subject analysis will be available in one place
Fulsome metadata is available to assist clients in the resource discovery process	Positions LAC to provide clients with more information on books; client targets the desired item more easily; greater client autonomy
If BookNet stores non-ONIX data, like covers and artwork, then LAC does not have to (LAC can “link in” to BookNet’s storage of this data)	Less LAC IT server space required to store and maintain non-ONIX data
Opens doors to future collaboration with book publishing community	Positions LAC for enhanced relationship with book publishing community for shared products and services

Annex 2:

Key Differences between the LAC and the Library of Congress Implementations of ONIX to MARC Conversion

Library and Archives Canada	Library of Congress
Converts approximately 140 fields from ONIX to MARC	Converts approximately 20 fields from ONIX to MARC
Stores records directly in AMICUS database in MARC 21 format	Stores records in a separate repository as text files
Conversions are performed on large files of records, all of which are then available through AMICUS.	LC MARC 21 conversions are performed on-demand and uploaded individually into LC database.
LAC conversion utility uses XSLT templates which are a commonly accepted means of converting XML-based documents. Using XML-based conversion tools allowed LAC to leverage XML tools developed for other applications.	LC uses an XSLT template developed internally. LC conversion utility does not use XML. ONIX encoding is stripped out and the data is stored as a text file.