The Role of Onix Data in Patron-Driven Acquisitions (PDA) Environment

Jackie Shieh
George Washington University
Libraries
2130 H Street NW
Washington, DC
01 202 994 6848
jshieh@gwu.edu

ABSTRACT
This paper explains the potential resource savings with the implementation of Onix data in a Web environment to facilitate patron-driven acquisitions (PDA) in an academic library. The George Washington University Libraries put forth a model that made use of the freely available publisher data in Onix XML format to facilitate in-house conversation with regards to PDA.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models, Search process, Selection process.* H3.5: Online Information Services – *data sharing, Web-based services*

General Terms
Management, Documentation, Performance, Design, Economics, Reliability, Experimentation, Standardization, Languages

Keywords
Onix, Metadata, Search and discovery, Library, Acquisitions, Patron-Driven Acquisitions, George Washington University Libraries, Data Transformation, Data Mashup

1. INTRODUCTION
Traditionally, the library's collections are resulted from arbitrary purchases—materials selected by subject specialists, collection development librarians. In most cases, some sort of data in MARC structure is required in order to initiate a purchase process. This need has thus created a variety of entrepreneurs that repurpose existing metadata with enhancements of classification and subject analyses for redistribution. [6] Much of library's funding resource has gone to support this endeavor.

However, researches have found that over 40% of book selected, purchased and cataloged in academic libraries were never circulated. Since early 2000's, the purchasing library materials began to shift to an "as needed basis" or a "consortial-based collection" which steered the library decision-makers to embark on projects that might secure alternate processes for library acquisitions, such as patron-driven acquisitions (PDA). [4] In the current economic environment, libraries are charged to be even more creative and innovative in experimenting and transforming the acquisition business for library materials. Many of the exciting experiments in the recent years have been those of the effort of repurposing data by employing existing helper application and open source tools.

Onix data (ONline Information eXchange) is a data exchange standard for publishers which contains 230 data elements representing "book, serial, and video product information in electronic form." [5] Even though, Onix has been established as a book industry standard for data exchange for B2B transactions facilitating ecommerce, the implementation among publishers, in particular the university presses, remains under performance. [8] This reality primary resulted from the lack of understanding of Onix and its potential usage beyond the publishing industry and inside the library community.

The early and continuing success of ecommerce from Amazon and other online services led publishers to realize that the revenue shortfall is caused by the absence of Onix data for their publications. As a result, a series of entrepreneurship began to create helper application that assists publishing industry to incorporate automatic Onix metadata generation to retailers and wholesalers at a push of a button. [9]

Since 2009 Onix data gained impressive momentum in the library community due to major efforts from the Library of Congress [11] and investment from Research Office at OCLC, Online Computer Library Center. [7] In 2010, multiple vendors' providing PDA service has been ever so present throughout the Charleston Conference programs. [4]

The George Washington University Libraries like many of its peer institutions faces similar challenge with regards to do more and faster with less and diminishing resources. The mandate of taking full advantage of technology to enrich user experience led us to investigate and explore Onix data as we consider its full potential of transforming the way library acquires and manages collection.

This pilot project that the George Washington University Libraries embarked on was to ingest the freely available publisher data in Onix format to facilitate library collection acquisitions process. The goal was to repurpose the data and make known the availability of new publications that are in-print, engage the library's users input in building the library's collections to meet the university's teaching curricula and research activities.[1] The

[1] A "library user" here refers to a person who possesses a valid library card.

just-in-time acquisitions weigh as important as the just-in-case collection.

2. METHODOLOGY and WORKFLOW

2.1 Data Ingest

The process begins with Onix data of 12814 titles representing three publishers: the University of California Press, Cambridge University Press, and Random House.[2] RSS feed is one of the mechanisms that publishers disseminate Onix data.[3] The data are coded following the Release 2.1, revision 02 (2004) standards.

Data files are located in an Apache server that runs SUSE Linux operating system. Figure 1 denotes the files structure for processing publisher's Onix data.

PATH		TYPE	DESCRIPTION
/srv/www/htdocs/		directory	this is the Apache root document directory
:....	wish/	directory	
:....	**index.html**	file	default web page displays hello message
:			calls wishWelcome.cgi
:....	data/	directory	
:....	filename.xml	file(s)	original onix xml file(s)
:....	**ONIXtransform.xslt**	file	XSLT stylesheet reads filename.xml creates onix2html.txt
:....	onix2html.txt	file	result of stylesheet's transform of onix xml
:....	**wishFormatter.pl**	file	Perl script that strips all newlines from onix2html.txt
:....	**wishFormatter.sh**	file	Shell script that calls Perl, then replaces ENDREC
:....	saxon.rows	file	transformed ONIX XML data as HTML rows and cells
:....	requests.txt	file	saved recommendations as pipe-delimited flat file
:			
:....	wishDisclaim.txt	file	Boilerplate text displayed on the thank you page
:....	wishStaffHelp.txt	file	Help file for staff
:....	titlebrowselist.txt	file	list of titles, used by the browse feature
:....	authorbrowselist.txt	file	list of authors, used by the browse feature
:....	sampleExcel.xls	file	sample Excel file of imported requests
:....	sampleAccess.accdb	file	sampe Access database file of imported requests
:....	onixexport.rptdesign	file	sample BIRT report design of imported requests

Figure 1. File Structure for Data Ingest

2.2 Data Transformation

Onix data contain very rich contents with respect to publisher's workflow. The Book Industry Study Group's (BISG) *Best Practices* for Onix data provides guidance for fields to include. [2] Bowker Data Services categorizes the grouping for ease of processing as follows: Core metadata (basic record), Descriptive metadata, Enriched components metadata and Specialty product metadata. [3] Figure 2 below shows which fields are the essential ones for presenting an information resource.

Core Metadata – Basic Record

Please send a basic bibliographic record containing the following data in order to ensure your titles are loaded to Bowker's products. Further details on each field can be found in the BISG Best Practices document link or the ONIX specifications from editeur.org as noted earlier in this document.

Description of Product:	*Binding/Format:*	*Price and Availability:*
Product Identifier *(ie: ISBN)*	Product Form	Publishing Status
Title Text	Product Form Detail *(if applicable)*	Publication Date *(yyyymmdd)*
Subtitle *(if applicable)*		Publisher & Imprint
Contributor *(include function)*		Supplier Name
Audience		Price *(include price type & currency)*
BISAC Subject Code		Discount Code

Figure 2. Bowker Core Metadata

A series of scripts in a Linux environment is in place to achieve the deliverables. Firstly, the ingested XML file gets transformed via Saxon in rows and cells. Perl, CGI, and Shell scripts interact with user inputs via a Web-interface that searches and recommends the data in HTML output. The output file has the EAN number, company and person information from the Header element, followed by title, author, biographical data, ISBN, imprint, extent, supplier, subject, subject code, description, series when applicable, and a thumbnail image file if available within a "product" element.[4]

2.3 Data Presentation

A Web-interface program with series of CGI scripts interacts with user as shown in Figure 3.

PATH		TYPE	DESCRIPTION
/srv/www/cgi-bin/		directory	
:....	wishWelcome.cgi	script	shows user a welcome page; calls wishLogin.cgi
:....	wishLogin.cgi	script	prompts user to login; calls wishSearchForm.cgi
:....	wishSearchForm.cgi	script	prompts user to enter selection criteria;
:....	wishRecommend.cgi	script	displays results of search; calls wishConfirm.cgi
:....	wishConfirm.cgi	script	displays selected titles; calls wishComplete
:....	wishComplete.cgi	script	stores selection; displays thank you message
:....	wishLoginStaff.cgi	script	prompts staff to login
:....	wishReview.cgi	script	staff interface for viewing requests
:....	wish/	directory	
:....	wishCategories.txt	file	contains options in category drop down list
:			used by wishSearchForm.cgi
:....	wish.cnf	file	configuration settings file, used by all scripts

Figure 3. File Structure for Data Processing

The Login page prompts user to enter an email address ending with gwu domain. This is to ensure requestor's legitimate association with the University.[5]

A search term is sent to query the data. When a match or matches are found, the result page prompts user to select from list to

[2] This number is the November feed from the three publishers representing, 8963, 1491 and 2360 respectively. Data is stored as flat file not in a relational database. If performance becomes an issue, data can be pre-filtered for specific type of materials.

[3] There are several models with regards to data feeds. Most commonly used are RSS, email notification, sftp, etc. The frequency of data source varies from publisher to publisher. Individual library needs to decide on a policy of how often the data must be refreshed.

[4] This link provides screen shots for the series of activities described for the pilot project, http://gwdroid.wrlc.org/JCDL2011/OnixGWU_Pilot.avi .

[5] User authentication can connect to the university's LDAP server.

recommend for purchase. Upon the user finalizing the recommended title list, the data is date-stamped and appended to an existing file awaiting a collection development librarian's review.

2.4 Recommendation Review

The output data is collected in a pipe delimited text file format for ease of import or extraction to different applications.[6] Using a Web interface for a staff member, a collection development librarian for a specific subject may choose to review a subject-based or the complete listing of recommendation from a given user or at a time-period.

```
DT|2011/01/19 12:04:47|J S|*****@gwu.edu|5354718091703O7|128.164.213.113|NO|1135|ITM|TI|Landscapes, Gender
DT|2011/01/19 12:04:47|J S|*****@gwu.edu|5354718091703O7|128.164.213.113|NO|3|ITM|TI|Looking at Greek Art|
DT|2011/01/19 12:04:47|J S|*****@gwu.edu|5354718091703O7|128.164.213.113|NO|540|ITM|TI|Local Knowledge and
DT|2011/01/19 12:04:47|J S|*****@gwu.edu|5354718091703O7|128.164.213.113|NO|685|ITM|TI|Archaic Greek Epigr
```

Figure 4. Output File

Each line in the file contains

Requestor information

Date time|user name|user email|user session id|user

IP address|

HTML line number **NO**|record id number|

Item detail **ITM**|**TI**|title|**AU**|author|**SU**|subject code|

Info from ONIX file header Onix EAN header|

Onix vendor header|onix from header

Dependent on the level of users' activities, the active output file can be archived on a weekly or monthly basis so that the collection development librarians are not overwhelmed with incoming requests. The source file is also refreshed on a regular schedule.

3. POTENTIAL ROLE IN WEB SCALE DISCOVERY

Many libraries either have begun or are considering adding a discovery layer on its online catalog, locally created digital contents, publicly available digital collections and subscription based databases hosted elsewhere. There are several discovery services that allow library to expose collections from different data sources through a uniform search interface to its constituents. [10] All discovery services focus on collections that the library already has or resources to which the library subscribes, including some that make use of a Z39.50 protocol or SRU for external data discovery focusing on structural metadata such as MARC.

Onix data on the other hand forecasts what the publisher has to offer. It is not the type of data source that appears to be as urgent as the content existing in the library's current collections. Nevertheless, it can be extremely valuable to a doctorate student who may not want to begin a research project if a topic is being prepared for publication.

[6] An example for this is the selected titles can be transformed into MARC records to upload to a library's catalog in order to interact with library's Acquisitions and Financial system.

In addition, the element of author's biographical affiliation often gets dropped out when converting to the library-centric metadata format, MARC since Onix data is not part of traditional library data source.

The George Washington University Libraries is in very good company in considering a Web scale discovery service to connect university research community to its vast information repository.

4. CONCLUSION

This pilot project explored the role of a metadata that is non-traditional to library operations and its potential in forging a closer collaborative relationship among users, libraries and resource providers, in this case, publishers. Many policy issues remain on the table for review.

Most academic, if not all, libraries have approval plans in place with major booksellers, vendors such as EBL, YBP, etc.[7] As more vendors take up the challenge of providing PDA service, the compatibility of data feed among library's online system (financial and bibliographic), and vendors' invoicing requires a closer and deeper investigation. Opportunities abound for library to be even more innovated in mashing up and serving data to meet the financial and economic challenges. From the limited experience of this pilot project, incorporating publisher Onix data as a supplementary data source in the overall research discovery adds a new horizon to the University's research and teaching community. With some ingenuity, perhaps, library users get to obtain a deeper and serendipitous discovery experience.

In a 2004 Australian Publishers conference, Cambridge University Press was quoted: "Onix has the potential to free up resources for better service… in our continuing pursuit of that perfect marketing match the right book for the right customer." [1] I couldn't agree more.

5. ACKNOWLEDGMENTS

This pilot project would not have been possible without major assistance from Michael Cummings, Library Systems Coordinator and Joshua Gomez, Digital Programmer Analyst at the George Washington University Libraries.

6. REFERENCES

[1] Australian Publishers Association. *Onix and TitlePage <Smart Data for Books>*. Slide #37. 2004. Book Industry Roadshow (2004) URL=http://publishers.asn.au/emplibrary/ONIX_2004.ppt

[2] Book Industry Study Group (BISG). 2005. *Product Metadata Best Practices for Data Sender*. Version 1.1. URL= http://www.bisg.org/docs/Best_Practices_Document.pdf

[3] Bowker Data Services. 2009. *Onix Data Submission Guide*. (10 August 2009) URL= http://www.bowker.com/products/DataSubmissionGuide_ONIX.pdf

[4] Charleston Conference. 2010. *Schedule At-A-Glance*. Presentations on the patron-drive acquisitions and collection

[7] Mostly of these PDA services focus on e-books.

development were across all levels. URL= http://www.katina.info/conference/downloads/2010AtAGlance_FINAL.pdf

[5] Editeur. 2005. *ONIX for Books: Product Information Message Overview and Data Elements*, Release 2.1. rev. 02. (July 2004) URL=http://www.editeur.org/files/ONIX%202.1/ONIX_Books_Documentation_2.1_rev.02+codes_Issue_11.zip

[6] Luther, Judy. 2009. *Streaming Book Metadata Workflow*. A Paper prepared for the National Information Standards and Organization (NISO) and OCLC, Inc. (June 30, 2009). URL= http://www.niso.org/publications/white_papers/StreamlineBookMetadataWorkflowWhitePaper.pdf

[7] OCLC Press Releases. 2009. *OCLC Now Offers Metadata Services for Publishers to Enhance Title Metadata for Use in the Supply Chain.* (14 October 2009) URL= http://www.oclc.org/news/releases/200954.htm

[8] OCLC Symposium. 2009. *Report on OCLC's Symposium for Publishers and Libraries.* In: *Symposium for Publishers and Librarians : Bringing Together Publishers and Librarians to Explore Metadata Needs and Practices*. (March 18-19, 2009) Dublin, OH. URL= http://www.oclc.org/publisher-symposium/summary/default.htm

[9] Onix Software for Publishers. 2010. URL=http://www.bookpublishingsoftware.com/web-onix.htm

[10] Vaughan, Jason. 2011. Web Scale Services. *Lib. Tech. Reports*. 47, 1 (Jan. 2011)

[11] Williamson, David. 2009. *Electronic CIP : The Cataloging in Publication Program*. Library of Congress. URL= http://cip.loc.gov/onixpro.html