SRELS Journal of Information Management
Vol. 42, No. 4, December 2005, Paper AN. p523-532.

ISSUES IN METADATA CROSSWALKS: A CASE STUDY OF QUALIFIED DUBLIN CORE AND ONIX

Renu Seth; A.R.D. Prasad*** and *Devika P. Madalli****

In the traditional library environment Cataloging in Publication (CIP) is very popular and the library staff can make use of CIP records to generate minimal cataloguing information. With the advent of Internet and the presence of book industry on the Net, the growing importance of metadata has resulted in ONIX (ONline Information eXchange). The benefits of crosswalk between ONIX and Dublin Core are too obvious to be ignored, as it saves the time of cataloguers in generating metadata in Dublin Core for the e-books or even printed books acquired by a library. This paper attempts to investigate the possibility of generating required metadata from available metadata formats, particularly the most popular Dublin Core (DC) from ONIX and presents mapping between ONIX and Qualified Dublin Core.

KEYWORDS/DESCRIPTORS: Cataloguing In Publication (CIP), Metadata, Online information exchange, ONIX, Dublin core, Qualified Dublin core, Mapping

1 INTRODUCTION

In the context of library networks major issues were focused around formats for bibliographic data exchange. It was necessary that the participating libraries and institutions use a common code/rules for bibliographic data and also agree upon a common record format so that the records from one could be harvested by another institution or union catalogues could be built with accessibility to all users of a network. A similar case is with metadata associated with digital objects in repositories and is an issue to be addressed to achieve interoperability across digital library collections. To reach the broadest community of information workers, metadata must be made available in accordance with a number of popular content metadata standards. As the number, size, and complexity of content metadata standards continues to grow, supplying the metadata for each standard becomes more and more repetitious, time consuming, and tedious [1].

* *Library Incharge*, Max Mueller Bhavan, New Delhi, INDIA. E-mail: *seth@delhi.goethe.org*

** *Associate Professor*, Documentation Research and Training Center, Indian Statistical Institute, Mysore Road, 8th Mile, Bangalore 560 059, INDIA. E-mail: *ard@drtc.isibang.ac.in*

*** *Lecturer*, Documentation Research and Training Center, Indian Statistical Institute, Mysore Road, 8th Mile, Bangalore 560 059, Bangalore, INDIA. E-mail: *devika@drtc.isibang.ac.in*

Creating metadata requires the knowledge of the standard and understanding of the content of each element unambiguously. With the many standards being introduced for every kind of repository and type of resources, the creators of digital objects are rather confused on what standard to use and data to furnish. One of the issues of the networked environment is that in an environment that is so dynamic and open to change, there is a greater and greater emphasis on standards [2]. But the paradox is the plethora of standards that have come with variations that defeat the purpose of a standard. However, it is important to achieve interoperability despite varied standards being followed for metadata. Ideally the metadata in one standard should be made accessible through crosswalks to another. Crosswalks evolved from the need for online information systems to cope with the metadata standards that have been developed in response to the recent onslaught of digital material, which presents concerns not addressed by standards developed for traditionally published work [3].

The purpose of this article is to lay the ground work for a crosswalk between ONIX and Dublin Core Metadata standard to facilitate exchange of bibliographic and product data between the library community and the book industry. Although the objectives and requirements of each community are different they have in common the requirement to accurately describe and locate documents. The expectation is that each community would benefit from the facility to reuse the others' data [4]. In the case ONIX, brought out by the publishing community, it could serve as a source from which metadata for digital libraries can be extracted, as it is quite elaborate with about 235 elements. ONIX gives the resources a better visibility and searchability because of the granularity followed. Again for most digital repositories, Dublin Core is adopted as a standard especially for exposing data to be harvested in an interoperable environment. A crosswalk from ONIX to Dublin Core would facilitate automatically elucidating metadata to enhance interoperability among digital library collections. This will be far less tedious than having to create the metadata records from scratch.

2 DUBLIN CORE

The Dublin Core Metadata Initiative (DCMI) *(www.dublincore.org)* is dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems.

2.1 Mission and Scope

The stated mission of DCMI is to make it easier to find resources using the Internet through the following activities [5]:

1. Developing metadata standards for discovery across domains,
2. Defining frameworks for the interoperation of metadata sets, and,
3. Facilitating the development of community- or disciplinary-specific metadata sets that are consistent with items 1 and 2

The range of activities of DCMI includes: Standards development and maintenance, Tools, services, and infrastructure, including the DCMI metadata registry and Educational outreach and community liaison. Ongoing efforts of DCMI participants include the collaborative development and continual refinement of metadata conventions based on research and feedback between DCMI Working Groups.

Dublin Core metadata provides card catalog-like definitions for defining the properties of objects for Web-based resource discovery systems. The Dublin Core is a set of eighteen generic metadata elements for discovering resources across a diversity of domains and languages. The core elements of Dublin core are -- Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights, Audience, rightsHolder and Provenance.

To meet specialized requirements, Dublin Core can be customized with additional elements or qualifiers. However, these refinements can compromise interoperability across applications [6]. DCMI recognizes two broad classes of qualifiers:

- *Element Refinement:* These qualifiers make the meaning of an element narrower or more specific. A refined element shares the meaning of the unqualified element, but with a more restricted scope. Example: element "Contributor" may be qualified by "Editor"
- *Encoding Scheme:* These qualifiers identify schemes that aid in the interpretation of an element value. These schemes include controlled vocabularies and formal notations or parsing rules. Example: element "Subject" may be qualified by "LCSH" implying that the subject keywords are according to LC subject headings.

2.2 Dublin Core in XML

DCMI has defined the elements and their qualifiers. But the Dublin Core elements have to be represented in a language to be interpreted by the web browsers. Increasingly XML is used to represent the metadata records. The main reason for this is that any element can be used to semantically represent information in XML as compared to the closed set of tags used by HTML. DC should be encoded *properties* as XML elements and *values* as the content of those elements. The following are the guidelines for representing DC elements in XML [7]:

- A *simple DC record* is made up of one or more *properties* and their associated *values*.
- Each *property* is an attribute of the *resource* being described.
- Each *property* must be one of the DC Metadata Element Set elements.
- *Properties* may be repeated.
- Each *value* is a literal string.
- Each literal string *value* may have an associated language (e.g. en-GB).

A sample record of DC in XML is as follows:

```
<record>
        <dc:title>introduction to Dublin Core</dc:title>
        <dc:creator>DCMI</dc:creator>
        <dc:identifier> http://purl.org/dc </dc:identifier>
 </record>
```

3 ONLINE INFORMATION EXCHANGE (ONIX)

ONIX stands for ONline Information eXchange [8]. It is an international standard for representing and communicating book industry product information in electronic form. It is a metadata standard developed by the publishing community as a standard means to exchange information about "book" product information electronically to wholesalers, retail booksellers, other publishers, and anyone else involved in the supply chain. The American Association of Publishers (AAP) developed ONIX during 1999 in conjunction with the major wholesalers, online retailers and book information services. It is intended to provide publishers a means of sharing product and supplier information usable on the Internet. ONIX was designed as a solution to two major problems [9]:

- The lack of consistency and standards in data exchange formats in use by book wholesalers and retailers and the need for a universal, international format in which all publishers could exchange information; and
- The need for richer book data online since there is no physical book for the potential buyers to pickup and pursue on the Internet.

ONIX provides an XML message format for exchanging information between systems, which may, internally, use different metadata systems. ONIX data elements have been defined for product information — message headers, reference and product numbers, authorship, subject, publisher, and the like — for books, and related media products. As a standard, ONIX is maintained primarily by EDItEUR, a membership organisation focused on standards for electronic commerce in the book and serials industries.

ONIX is a rich metadata scheme that is comprised of 235 elements of information that fall into 24 categories. Each element is defined to ensure consistent use. Some elements are required (such as ISBN, author, title) and some are optional (such as book reviews, cover image). The element set includes descriptive, administrative and structural metadata elements. The level of granularity of information is finer than that has been developed by MARC/AACR2. Although most of the descriptive metadata elements map to MARC, many of the administrative and structural elements in ONIX do not have equivalence in MARC [9].

ONIX uses a series of data elements that describe book information. The data elements are simple identifiers enclosed in angular brackets. For example, the tag <a01> is used to indicate an ISBN, while <d01> identifies the title. The ONIX documentation type definition (DTD) contains in its entirety over 230 data elements and composite elements, organized into 38 groups: 25 of which relate to product records, 6 to main series records and 7 to sub-series records. The data elements listed below are those, which have been identified as comprising the key elements in a product description and are categorized as follows [8]:

- M Mandatory (i.e. required for all products and measured accordingly);
- R Required under specified conditions (i.e. required for all relevant products or in all relevant situations, and measured by reference to the presence of data in other fields); and
- E Expected to be used when applicable. Not capable of strict measurement, but repeated failure to provide expected elements might disqualify publishers from accreditation.

4 METHODOLOGY FOR THE CROSSWALK

Crosswalks can be implemented in many areas of work. Some of the probable scenarios are given below:

1) Z39.50 server

2) OAI-PMH implementation/ Digital Libraries

Any repository's metadata can be exposed using either Z39.50 protocol or OAI-PMH (Open Archives Initiative-Protocol for Metadata Harvesting).

4.1 Z39.50 Protocol

The best-known open source software for Z39.50 protocol is the Zebra Server of Indexdata *(http://www.indexdata.dk)*. The Zebra server accepts records either in ISO-2709 format or XML format, though in case of both Dublin Core and ONIX, it is ideal to use records in XML format. In fact, the YAZ toolkit from Indexdata can be used to convert ISO-2709 format records into XML. The most common practice in generating metadata records is to create an XML file for each record, though creating one XML file to have all the records is not quite uncommon. Zebra offers good documentation on creating index to the metadata of your repository, using a Z39.50 client one can search, retrieve and download metadata. Zebra helps defining crosswalks and provides the data in any schema (DC or ONIX or MARCXML) and in any format (XML, ISO-2709, HTML).

OCLC's SRU/SRW, based on ZING (Z39.50 Next Generation) is yet another software, which attempts to get best of Z39.50 and HTTP (Hyper Text Transfer Protocol). SRU/SRW accepts records in XML format. SRU/SRW of OCLC is quite flexible, where metadata of all records can be in one file or in separate files. It also handles metadata stored in various databases files e.g. it can handle DSpace records directly from PostgreSQL database.

A simple program in Perl or any programming language can crosswalk records from ONIX and Qualified Dublin Core.

4.2 OAI-PMH

OAI-PMH is becoming popular with the popularity of Open access to information movement. Digital repositories and open access journals are adopting OAI-PMH to make their metadata available to search engines and harvesters. Many digital repository software have inbuilt mechanism to expose metadata using OAI-PMH protocol, e.g. DSpace and EPrints. In fact, DSpace uses OCLC OAICat for exposing metadata. There are quite a good number of OAI-PMH data provider software, which can handle various metadata formats. *http://www.oaiforum.org* enlists data provider software. A few of them, which we tried, are mentioned below:

1) DLESE (can be used both as service provider and data provider) (*http://www.dlese.org/oai/index.jsp*);
2) OAICat (exclusively to act a data provider) (*http://www.oclc.org/research/software/oai/cat.htm*); and
3) PKP-Harvester (though it mainly used for harvesting metadata as service provider, it has provision to make it a data provider) (*http://pkp.sfu.ca/pkp-harvester/*).

5 CROSSWALK USING XSLT

Using XSL Transformation, one can easily define rules for mapping from one metadata format to another metadata format. XML is good for data import and export but for exchanging XML data from one standard to other standard and mapping of the elements between two standards, crosswalks are required. XSLT is an easy way to transform data from one metadata standard to other. Several crosswalk maps can be found over Internet, which could be codified in XSLT statements for easy transformation. [10]

6 CROSSWALK IN DSPACE

DSpace by default uses Qualified Dublin core and exposes metadata in Unqualified Dublin core format for OAI-PMH purposes. As DSpace uses OCLC's OAIcat, it can be easily extended to other metadata schemes by developing java programs. The recent version of DSpace, version 1.2.2 allows users to define their own metadata formats using *inputforms.xml.* One can add metadata elements in the database tables and define the various input worksheets. However, to expose the metadata or perform crosswalks one should write java programs, which should not be too difficult. A detailed discussion of dealing with multiple metadata formats was discussed in a paper presented at the DSpace User meet [11].

7 MAPPING BETWEEN ONIX AND QUALIFIED DUBLIN CORE

ONIX has a huge set of elements and some of them are meant for business purposes. We may broadly categorize metadata elements to intrinsic and extrinsic elements. Here, we mean intrinsic elements are the elements that describe a document and its properties and are inseparable and inherent to the item. Whereas the extrinsic elements are the ones that are created for different purposes by the organization creating the metadata. For example, in case of MARC the accession number is not intrinsic to the documents being processed and the value varies from library to library. In case of crosswalk between ONIX and Dublin Core one may consider only the intrinsic elements and leave the extrinsic elements, as they mostly deal with information regarding sales workflow.

Table: Mapping ONIX elements *(selected)* to DC Qualified set

Onix Tag*	Reference Name**	Onix Element***	DC Elements	DC Qualifiers
	<Contributor>	Contributor	Contributor	
<b034>	<SequenceNumber>	Contributor Sequence Number	Contributor	
<b035>	<ContributorRole>	Contributor role code	Contributor	

Onix Tag*	Reference Name**	Onix Element***	DC Elements	DC Qualifiers
<b036>	<PersonName>	Person name	Contributor	
<b046>	<Affiliation>	Affiliation	Contributor	
<b047>	<CorporateName>	Corporate contributor name	Contributor	
<b044>	<BiographicalNote>	Biographical note	Contributor	
	<Title>	Title	Title	
<b032>	<TranslationOfTitle>	Translation of title	Title	alternative
<b203>	<TitleText>	Title text	Title	
<b029>	<Subtitle>	Subtitle	Title	
<b036>	<PersonName>	Person name	Creator	
<b047>	<CorporateName>	Corporate contributor name	Creator	
<b052>	<ConferenceName>	Conference name	Creator	
<b003>	<PublicationDate>	Publication date	Date	Scheme (–ISO-8601/ W3C-DTF)
<b087>	<CopyrightYear>	Copyright year	Date	dateCopyrighted
<b081>	<PublisherName>	Publisher name	Publisher	
<d101>	<MainDescription>	Main Description	Description	
	<ContentItem>	Content items composite	Description	TOC
<d100>	<Annotation>	Annotation	Description	abstract
<b219>	<ExtentValue>	Extent value	Format	extent
<b220>	<ExtentUnit>	Extent unit	Format	
<b061>	<NumberOfPages>	Number of pages	Format	
<b218>	<ExtentType>	Extent type code	Format	Scheme (mime type. IMT)
<b125>	<NumberOfIllustrations>	Number of illustrations	Format	
<b063>	<MapScale>	Map scale	Format	
<b216>	<EpubFormatDescription>	Epublishing format description	Format	
<c094>	<Measurement>	Measurement	Format	
<c258>	<Dimensions>	Dimension statement	Format	
<b004>	<ISBN>	ISBN	Identifier	
<b005>	<EAN13>	EAN-13 number	Identifier	
<b007>	<PublisherProductNo>	Publisher's product number	Identifier	
<b008>	<ISMN>	ISMN	Identifier	
<b009>	<DOI>	DOI	Identifier	
<b059>	<LanguageOfText>	Language of text	Language	
<b060>	<OriginalLanguage>	Original language	Language	
<b252>	<LanguageCode>	Language code	Language	ISO639-2; RFC3066
<b018>	<TitleOfSeries>	Series title	Relation	IsPartOf
<b207>	<AudienceDescription>	Audience description	Audience	
<b073>	<AudienceCode>	Audience code	Audience	

Onix Tag*	Reference Name**	Onix Element***	DC Elements	DC Qualifiers
<b189>	<USSchoolGrades>	US School Grade(s)	Audience	EducationLevel
<b064>	<BASICMainSubject>	BASIC main subject category	Subject	
<b070>	<SubjectHeadingText>	Subject heading text	Subject	
<b067>	<SubjectSchemeIdentifier>	Additional subject scheme Identifier	Subject	
<b171>	<SubjectSchemeName>		Subject	DDC, LCSH
<b071>	<CorporateBodyAsSubject>		Subject	
<b072>	<PlaceAsSubject>		Subject	
	<PersonAsSubject>		Subject	

****Onix Tag Numbers*** are numbers assigned to ONIX Reference Name.
*****ONIX Reference Names*** are standard names assigned to each ONIX Element.
******ONIX Elements*** are the qualifications of products produced by the publishers.

8 CONCLUSION

Unfortunately, the specification of a crosswalk is a difficult and error-prone task requiring in-depth knowledge and specialized expertise in the associated metadata standards. Obtaining the expertise to develop a crosswalk is particularly problematic because the metadata standards themselves are often developed independently, and specified differently using specialized terminology, methods and processes. Furthermore, maintaining the crosswalk as the metadata standards change becomes even more problematic due to the need to sustain a historical perspective and ongoing expertise in the associated standards [1].

As there cannot be one-to-one relation between the element sets of various metadata schemes, crosswalk across metadata schemes would always be lossy. In theory, one-to-one, many-to-one mapping can be done without loss, whereas one-to-none, none-to-one, one-to-many would result in loss or incorrect transformations. In case of MARC records, retroconversion (crosswalk) though possible, it was never without loss of data. However, in case of MARC the structure of the descriptive elements of various MARC was main reason for loss of data. The structure of elements between MARCs could not be mapped exactly. However, in case of metadata schemes, the descriptive elements of different schemes greatly vary resulting in many one-to-none or none-to-one relations. Perhaps, a satisfactory solution for the harvester is to use selective harvesting by collection and use corresponding metadata format of that collection. This approach is only relatively satisfactory between the contending metadata formats. For example, if there are more than one schemes for electronic theses and dissertations, the different metadata formats would be describing the same type of digital item, though differing slightly in their description and the element set they use. Alternatively, one has to accept Dublin Core as the lowest common denominator.

REFERENCES

1. Pierre (Margaret St) and LaPlant (William P. Jr). *Issues in Crosswalking Content Metadata Standards*. http://www.niso.org/press/whitepapers/crsswalk.html
2. Hodge (Gail M). Best Practices for Digital Archiving: An Information Life Cycle Approach. *D-Lib Magazine*. Vol. 6(1); January 2000. http://www.dlib.org/dlib/january00/ 01hodge.html
3. Godby (Carol Jean); Young (Jeffrey A) and Childress (Eric). Repository of Metadata Crosswalks. *D-Lib Magazine*. Vol. 10(12); December 2004. http://www.dlib.org /dlib/december04/godby/12godby.html
4. Danskin (Alan). *Report on an ONIX UNIMARC crosswalk. http://www.bic.org.uk/ reporton.doc*
5. *About the Initiative*. http://dublincore.org/about/
6. Baker (Thomas). Dublin Core in multiple languages: Esperanto, interlingua, or pidgin? **In** *Proceedings of the International Symposium on Research, Development and Practice in Digital Libraries*. 1997. University of Library and Information Science, Tsukuba, Japan. http://www.DL.ulis.ac.jp/ISDL97/proceedings/thomas/thomas.html
7. Powell (Andy) and Johnston (Pete). *Guidelines for implementing Dublin Core in XML*. 2nd April 2003. http://dublincore.org/documents/dc-xml-guidelines/
8. *ONIX for Books*. http://www.editeur.org/onix.html
9. *ALCTS Taskforce on ONIX*. http://www.libraries.psu.edu/tas/jca/ccda/tf-onix2.html
10. Tripathi (Aditya). Metadata Crosswalks with MarcEdit using XSLT. **In** *DRTC Workshop on Semantic Web. 8th - 10th December 2003*. DRTC, Bangalore. http://hdl.handle.net/1849/127
11. Prasad (A R D). Using Multiple Metadata formats in DSpace. **In** *DSpace User Meet. 6-8th July 2005*. University of Cambridge, Cambridge, UK.
12. *Booknet Canada, Bibliographic Data: ONIX*. http://www.booknetcanada.ca/booknet/biblio_ data.shtml
13. *ONIX Implementation Tutorial – Introduction*. http://abiblion.com/onixtutorial/
14. Brand (Amy); Daly (Frank) and Meyers (Barbara). *A guide for Publishers: Metadata demystified*. http://www.niso.org/standards/resources/Metadata_Demystified.pdf
15. Baker (Thomas) and Dekkers (Makx). *Identifying Metadata Elements with URIs*. http://www.dlib.org/dlib/july03/baker/07baker.html
16. Lanzinger (Susan S). *Digital Preservation and Metadata: History, Theory, Practice*. Englewood, CO: Libraries Unlimited, 2001.
17. *ALCTS report: Task Force on ONIX International*. http://www.libraries.psu.edu/tas/jca/ccda/ tf-onix2.html