ONIX for Books, MARC and Digital Preservation of eBooks

Assignment 4: Completed Research Paper

Yu Ting Lin
LIBR 281-11
Dr. Mary K. Bolin
Introduction

Emergence of digital technology brings convenience and efficiency to many aspects of information creation, access and distribution in the contemporary age that was unimaginable before. Available in digital formats, information beyond print copies can easily be provided to multiple users at distant locations simultaneously, allowing greater freedom on information access (Garrett & Waters, 1996). The relatively easy reproduction and modification process of digital materials also makes possible the diverse information utilizations, alleviating the traditional conflict between preservation and creation. Furthermore, technology enables novel ways to organize, retrieve and display information, generating an information experience that is more efficient and user-friendly. Advantages of digital technology and materials make this relatively new format rapidly pick up an important role among users and become one of the popular trends in the contemporary information community.

However, benefits of digital information come with new challenges and problematic complications for information preservation. Content of digital materials relies on software and hardware for storage of data and interpretation into readable information (Hedstrom and Montgomery, 1998). Longevity of digital information thus depends not only on the preservation of the actual information documents, but also on the availability of the corresponding software and hardware technologies. The rapid change of technology makes it difficult to prevent information recording and interpreting medium from being obsolete, risking information loss along with the technologies. Moreover, information technology lowers the barrier for information edition and publication, resulting in an overflow of assorted editions of the same materials and self-
published documents. Convenience of information creation expands the diversity of knowledge but also make preservation more difficult. If we don’t know what’s out there, how can we preserve it?

Though with a late start, books follow the trend of digitalization. By December 2010, eBooks already made up about 10% of the general book sales, and the percentage is expected to rise even higher in the next five years (Kirchhoff, 2011; Tracy, 2008). EBooks has shown a clear pattern of progression into one of the most utilized formats for information consumption. Meanwhile, urgency for eBook preservation becomes a widely recognized topic among scholars and information organizations. In response the format’s unique features, there is an increasing consensus for the need to establish structured policies and activities specifically for eBooks to “ensure the enduring usability, authenticity, discoverability and accessibility of content over the very long term” (Kirchhoff, 2011, page 1).

**Statement of the Problem**

EBooks essentially shares many similarities with other types of digital materials, and thus, it also shares the challenges in information preservation. Electronic books’ unique characteristics further create obstacles that are specific to eBooks, such as unclear preservation responsibility due to eBooks’ license-based ownership model (Kirchhoff and Morrissey, 2014). For many of the already identified issues, until further clarification is settled, current preservation effort depends on voluntary collaborations among publishers, vendors, libraries and possibly third party preservation service organizations.
Nowadays, eBooks preservation relies on MARC records, which are mostly provided to libraries by publishers and vendors (Kirchhoff and Morrissey, 2014). Majority of publisher provided MARC records are originated in non-MARC formats, such as in ONIX for Books, a popular record management system among publishers and vendors. These non-MARC format metadata then are converted into MARC records through utilization of assorted mapping techniques. However, as many scholars point out, conversion often lead to inconsistency of record quality, which might hinder accessibility and preservation (Traill, 2013).

This paper focuses on examination of the mapping between ONIX for Books metadata and MARC records. Official metadata creation guidelines and conversion procedures are the two primary emphases for comparison of similarities and differences of the two systems. The paper aims to help publishers and vendors understand effectiveness of such mapping techniques. On the other hand, the paper also wants to help libraries and preservation organizations identify error rate for publisher provided MARC records. Through detailed analysis, the paper seeks to reveal elements that contribute to success and/or failure of ONIX for Books metadata in support of MARC records, and determine whether publishers’ use of ONIX for Books is a cause of poor record quality and thus contribution to difficulty of digital preservation for eBooks.

Research Question

How well does ONIX for Books metadata convert into MARC records? Does application of ONIX for Books help or prevent digital preservation of eBooks?
Methodology

The study’s primary goal is to reveal level of achievement in the conversion of ONIX for Books metadata to MARC records; therefore examination primarily focused on the design of the crosswalk. Mapping guidelines and supporting documents, such as ONIX for Books best practices and specifications for subject fields, were compared and analyzed. All documents used in the examination were retrieved from one of the following websites: ONIX for Books website, MARC guideline website and OCLC official website. Discussion is based on content analysis of guidelines and documents. Through close examination and comparison of data, the paper hopes to identify elements that help or prevent ONIX for Books metadata be successfully converted into MARC records, and determine the level of effectiveness of the conversion.

Literature Review

Digital Information and The Library

Emergence and evolving information technology has brought the information community to the next stage of history with the prosperity of digital materials. The conversion from print to digital creates a new model of information access, creation and distribution that is faster, cheaper and easier (Riley-Reid, 2015). The trend of digitalization brings an opportunity for libraries, but at the same time, also highlights value of libraries' traditional roles. Information technology and digital material lower the barrier of information distribution, which however, leads to overflow of redundant information and inefficient organization (Neal, 2015). Libraries’ expertise in selecting, acquiring and synthesizing information is even more important now to help patrons figure out ways to navigate and retrieve
needed information from the massive and growing universe of information. Moreover, though information technology help reduce gap of accessibility, technology itself actually creates another type of barrier. Libraries’ provision of technological support and technology literacy assistance again helps lift the barrier to enable information access among greater population (Riley-Reid, 2015). Last but not least, the increased amount of information available raises the necessity for a more structured preservation strategy. As a result, though experiencing shifts in in formats of information handled and process engaged, the library continues serving the traditional roles of information provider, gatekeeper and preserver.

**Digital Information and Preservation**

Digital information enables the library to approach the topic of information organization and access from a different angel. The advantage of digital materials is clear; however, digitalization of the information world is merely the beginning of new challenges. In contrast with traditional medium, digital information usually requires external equipment and software to interpret, read and understand (Preserving digital information). As Jeff Rothenberg explained in his 1999 publication, "the bits in each document file are meaningful only to the program that creates the file...a document file is not a document in its own right: it merely describes a document that comes into existence when the file is run by the pogrom that created it" (page 9-10). The unique characteristics of digital material thus raised the new need to not only preserve the content but also the corresponding hardware and software.
Realizing the digital information’s dependency on the physical media and software, early attempt focused on the preservation of technologies necessary for digital data decoding and retrieval (Preserving digital information). However, “digital media can be fragile and have limited shelf live” (Preserving digital information). Additionally, information technology’s rapid replace rate increased the difficulty to ensure that all versions of hardware and software necessary for reading collected data are preserved. Technological obsolescence became the first barrier in ensuring longevity of digital information. The succeeding attempt to transfer digital information by constantly adapting files into formats readable by up-to-date technologies also faced critical issues. Problem of loss of information during the transition process and possibility of non-compatibility between older formats and newer formats makes the process difficult (Rothenbern, 1999). Moreover, as for preservation of the hardware and software, the constant change of information technology makes it hard to ensure up-to-date formats. Even after overcoming the two major issues, the cost can be unbearable.

The struggle of digital preservation rang the bell among academic scholars; however, the urgency of digital preservation was undermined at first in the mainstream community. As a 1998 survey conducted by Research Library Group revealed, despite that two-third of the survey institutes recognized the importance of information preservation and assumed the preservation responsibility, only half of them were actively engaging in digital preservation, and even fewer had developed structured preservation policies (Hedstrom & Montgomery, 1998). In the past decades, as utilization of digital information boomed, the need of digital
preservation was brought to attention again. Many organizations, such as Library of Australia, the British Cedars Project, OCLC and the Research Libraries Group, began to invest in projects for development of standardized metadata schemes to ensure preservation of digital materials (National Information Standards Organization, 2004). Preservation systems and institutions, such as Portico, LOCKSS and CLOCKSS contributed to current success in the field (Kirchhoff, 2011). Today, though digital preservation’s challenges still exist, more and more people and organizations recognize the significance of the issue and work together towards a solution. Unlike in the late 20th century, this time, the community takes an active approach to protect its valuable digital materials.

**E-Books, Preservation and Issues**

EBooks was initially defined as “the presentation of electronic files on digital displays, whether that content would have been presented in non-digital form as a book, a magazine, a newspaper, or a catalogue” (Kirchhoff & Morrissey, 2014, page 4). However, as born digital electronic materials increase, definition of eBooks expands to include “e-only monographs, monographs appearing both in print and online, digitized print titles, and continuously updated reference databases” (Tracy, 2008, page 40). Though eBooks currently only accounts for a small portion of most of the library’s existing collections, Tracy’s survey among publishers and library managers both indicate trend of significant growth in the next few years.

As an emerging format of digital information, eBooks faces the same challenge of ill-defined preservation responsibilities and strategies as when digital information first utilized. In fact, the unique license-access model makes the situation even worse.
the library purchases eBooks, it does not permanently own the books as in the procurement of traditional print copies (Kelly, 2014). The library only gains temporary access to the information during the contract period mostly through either the publishers’ technology or servers (Kirchhoff, 2011). The “non-ownership model” limits the library’s ability to ensure long-term access to the content, as well as ability to preserve the hardware and software necessary for decoding the data (Kelly, 2014). On the other hand, though publishers show concern for preservation of eBook content, at the time of Tracy’s study in 2008, none of the six publishers interviewed already established working solutions for eBooks’ preservation.

In the current stage, eBook preservation and information communication follows the existing model of MARC metadata records. A MARC record is a machine-readable cataloging record commonly used among library systems (Library of Congress, 2009). Record may contain description of the item, subject headings and classification numbers. This established standard provides a consistent and reliable format of bibliographical data, enabling smoother data transition among different systems and thus supporting a stronger network of data communication and accessibility. Quality of MARC record thus determines the following success of information transition, retrieval and preservation.

However, even pointed out in the MARC guideline, inconsistent MARC data sometimes might hinder the full capacity of MARC records. In fact, such problem often exists on metadata originated in non-MARC format. The XML-based ONIX for Books records are one of the most popular non-MARC metadata formats, especially among publishers and vendors (Halverstadt & Kall, 2013). Similar to MARC, ONIX for Books is an “international standard for representing and communicating book industry product
information in electronic form” (Editeur, “Overview”). Records contain rich product information, including promotional and territorial rights, prices and other information specific to the book industry (Polanka, 2011). ONIX for Books enables effective communication in the supply chain, efficient internal information updates and smooth data transition between different systems (Editeur, “Overview”).

The high popularity of ONIX for Books among book industry and growing need for solutions to convert data between ONIX for Books and MARC leads to OCLC’s new services for publishers. Metadata for Publishers aims to help publishers translate ONIX files into MARC records, and at the same time, enrich the original ONIX file by retrieving data from the MARC records (OCLC). Through the service, OCLC hopes to help publishers decrease investment in MARC cataloging and increase quality of metadata, resulting in a win-win situation for both publishers and the library.

Success of OCLC’s services is yet to be examined, but quality of MARC records received from publishers continues to be a primary concern among libraries. Traill’s 2013 study of record quality for vendor-provided records reveals, “All of the eighty-nine record sets (each record set includes 100-1,000 records) exhibited at least one error. About one-fifth displayed critical errors, while the vast majority of sets displayed at least one access error” (page 216). Halverstadt and Kall also raise concerns for common errors, such as name headings, special characters and subject headings, for conversion of non-MARC records to MARC records (2013). Such a high error-rate requires additional staff for record clean up, which might result in extra cost or delay in material access. Poor quality thus seriously diminishes the advantages of utilizing MARC records. In addition to the known difficulties for digital preservation and the problems emerged from
eBook’s unique business model, quality issue thus is another important challenge that needs to be overcome for future success in eBooks’ access, retrieval and preservation.

Results

Effectiveness of ONIX for Books metadata's conversion into MARC records was evaluated in two parts. The first part examines similarities and differences of the overall structures of ONIX for Books (will be referred as “ONIX” in the following paragraphs) and MARC. The second part then focuses on comparison of individual fields of the two standards. Primary documents that are analyzed includes: “ONIX for Books Product Information Format Specification,” “ONIX for Books Implementation and Best Practice Guide,” “ONIX for Books Codelists Issue 29,” “ONIX 3.0 for Books–MARC 21 Mapping (Crosswalk)” and “MARC 21 Format for Bibliographic Data.”

Overall Structure

ONIX for Books is a XML based standard that was created to enable communication of product metadata among publishers, venders, retailers and more. An ONIX message usually consists of four components: the start of message, the message header, the body of the message and the end of message. Information contains in the four parts is described by XML tags, each of which stands for a field defining a specific characteristics of the message or the product embedded in the message. Alike MARC, ONIX tags that describe similar characteristics are categorized into groups. However, these two standards present their elements and groups quite differently. While MARC standard groups all related subfields into one primary field and separated each by punctuations marks, ONIX presents related tags in
a block format. One ONIX XML tag only reflects a single characteristic and every tag starts from a new line. Data tags that are for specific information then are nested within the parent composite tags, enclosing within the opening and ending tags of the parent tags.

For example:
In MARC record, one line of field 245 Title Statement includes several parts of information: entry added, non-filing characters, title and author.
245 10 Webster's New World Thesaurus / $c Charlton Laird

On the other hand, ONIX record presents information in block format, with each line describing a unique characteristic.
<TitleElement>
   <SequenceNumber>2</SequenceNumber>
   <TitleElementLevel>01</TitleElementLevel>
   <PartNumber>Book 3</PartNumber>
</TitleElement>

The format differences thus influence the overall element order and structures of records. The Crosswalk provides a simple solution for the differences. Many of the ONIX tags correspond to not a MARC field number, but more specifically the subfield or indicator of a MARC field. For instance, <NumberOfCopies> in ONIX corresponds to MARC Field 980 subfield g. Some more complicated tags may even be directed to different MARC subfields based on the value selected. For example, if the <TitleType> value is 02, then the <TitlePrefix> value should be converted into MARC field 222 2nd indicator to reflect length of title prefix; if <TitleType> value is 06, then the <TitlePrefix> value should be translated into MARC field 242 2nd indicator to reflect length to the title prefix. The detailed instruction listed in the Crosswalk provides a clear mapping between ONIX tags and MARC fields, subfields and indicators, and effectively resolve the issue of format differences.
In addition to format differences, content of ONIX metadata and MARC records also shows great differences. ONIX for Books is designed primarily to fulfill needs of publishers and others in the industry; thus it includes many fields that present information that is useful for this target user group, such as price, availability and supply. Furthermore, “to cover the widest possible range of needs, it therefore includes many elements which are specialized to particular forms of publishing or particular markets” (“ONIX FAQs”). On the other hand, MARC standard is designed more towards needs of libraries, focusing on bibliographic, authority and classification information of products. Number-wise, ONIX includes much more distinct fields than MARC does. However, this information coverage differences in fact should not affect quality of MARC records. The Crosswalk again provides clear instructions on transition between ONIX and MARC by specifying what to map and what not to map. ONIX tags that are not relevant in MARC records, such as <ProductAvailability>, <PriceType>, <DiscountCoded> and <CitedContent>, are not mapped into MARC data and thus not included in the resulting MARC records. As a result, irrelevant data would be excluded during the conversion, leaving only valid information for MARC records.

Employment of controlled vocabulary is another point of differences. Both standards utilize controlled vocabularies as means to achieve systematic organization of values, but their applications are different. ONIX applies controlled vocabularies for values of more than 100 tags, ranging from audience and file type to publishing roles and sales rights. All tags with restricted values have their individual corresponding “codelists” to define possible values, usually in forms of
two to three digit combinations of letters and numbers. MARC standard, on the other hand, applies controlled vocabularies mostly in the form of indicators. It also has a few control subfields with set values and a few fields that utilizes controlled vocabulary lists. In comparison, MARC has a lot more free text fields than ONIX does. ONIX’s more restricted value for product description actually makes the transition between different systems easier. In many cases, ONIX’s limited number of answers makes it easier for the Crosswalk to provide direct mapping between each value and corresponding action for MARC field transition. It also helps avoid minor problems such as spelling errors and format inconsistency.

However, though the Crosswalk provides thorough instruction for mapping ONIX values into MARC fields, there are still circumstances that cannot be absolutely defined and require additional determination and correction. As Library of Congress specifies in its instruction, “in some cases, interpretation and special processing would be required of the content of an ONIX data element to render data compatible with the content of the corresponding MARC 21 data element. Often a choice between more than one potential MARC 21 data element is required.” (Library of Congress, 2005, “II. ONIX to MARC 21 Mapping Table”). For instance, values for <ProductForm> may be interpreted into 10 distinct MARC fields/ subfields; thus, judgment of what MARC fields to include and what to neglect is important. Other fields, such as <ProductContentType> and <ProductFormDetails>, also have this issue of “homonymy.”
Individual Fields

In addition to comparison of the two standards’ overall structures, three fields—Title, Author and Subjects—were also carefully examined to evaluate similarities and differences of the two standards, and thus their compatibility.

ONIX product title information is primarily included in the <TitleDetail> composite, which aims to present title information in a structured fashion. Several tags are nested under the field, including <TitleType>, <TitleElement>, <TitleText> and more. <TitleDetail> is repeatable for annotation of different types of title, such as former title and original language title. Though MARC standard’s Title Statement Field 245, is not repeatable as <TitleDetail>, MARC metadata incorporates other fields, such as Field 222 Key Title, 246 Varying Form of Titles and 247 Former Title, to catch information of variants of titles. The Crosswalk provides straightforward instruction on converting information in the <TitleDetail> composite into various MARC fields and subfields. Additionally, both ONIX and MARC stress the importance to input title information as is on the product title page.

Similar to <TitleDetail> for product title information, authorship related information is all listed in the repeatable <Contributor> composite. <Contributor> also has a variety of data tags to ensure include of important elements regarding authorship, and each of which is carefully mapped to a corresponding MARC field. However, unlike product title where ONIX and MARC share the same content formatting, the two standards’ dissimilar instructions for author name entries cause data conversion difficult. ONIX <Contributor> composite requires “carry the name of the person or corporate contributor in the form it appears on the product” (“ONIX
Best Practices”), while MARC requires author name to be in the inverted form. Though ONIX also includes a tag for <PersonNameInverted>, but the Crosswalk does not offer a clear instruction on when to map <PersonName> and when to use <PersonNameInverted> for MARC field 100$a and 245$c. Instead, it directs both tags for the two author fields with no annotation to transform names into the inverted form. Incorrect name order might lead to confusion of the author’s first name and last name. Additionally, MARC field 100 is usually used to include the authorized form of authors’ names. The mapping from ONIX tags to field 100 not necessarily reflects such authority record, which again might cause confusion and inaccessibility.

Subject is another field that reveals difficulty of compatibility between the two standards. ONIX defines Subject as to “describe what or who the product is about independent of the of the physical or digital nature of the product” (“ONIX Best Practices”), and supports more than 100 different subject schemes for this field. ONIX encourages users to apply any schemes that are relevant. The Crosswalk attempts to capture all aspects of the field by establishing a referencing worksheet for the subject fields and corresponding MARC fields. However, instruction is only provided for values of about 15 subject schemes, neglecting possible use of other fields. Lack of instruction not only might lead then lead to confusion and data loss.

The overall structure and the three individual fields examined demonstrate a wide range of levels of similarities and differences between the two standards. The next section will continue the discussion on how these characteristics affect success of digital preservation of eBooks.
Discussion

ONIX for Books (will be referred as ONIX) and MARC present two record management systems with two quite distinct sets of guidelines. The overall structures of the two systems are compared in three different aspects: information grouping and presentation, types of information included and use of controlled vocabulary. Though all three categories demonstrate fundamental differences between ONIX and MARC, as discussed above, these information gaps can easily be bridged with instructions provided in the Crosswalk. The mapping guideline seems to provide an effective standardized process of data conversion, which can lead to creation of consistent quality.

However, we still need to keep in mind that conversion of certain fields requires professional judgments. The Crosswalk does not provide an absolute instruction for these fields. Rather, it leaves an open-ended space for flexibility of information inputs, originations and formats. Such flexibility can be beneficial in some cases, but in others, might lead to increase in error rate, diminishing the value of a standardized conversion procedure.

In comparison to the examination of ONIX and MARC from the overall structure approach, investigation of individual fields yield an interesting set of results with a wide range of compatibility. Again, three core fields were selected to compare similarities and differences of the two systems and how these elements affect the overall success of data conversation. Among the three, Title is the only field that demonstrates high potential of compatibility between the two systems. Authorship and Subject on the other hand presents complications that might lead to critical errors in MARC records. Differences in
data format, in the Authorship example, the spelling and presentation of author names, very like would lead to error and inconsistency during the conversion, and further result in inaccessibility of the end record. The Subject field reveals a similar issue. Inequality of information captured for the Subject field in ONIX and MARC makes it difficult to translate relevant data from one to the other, and thus, might lead to incomplete or misleading information.

The structure and guidelines of ONIX and MARC each well reflects its goals as a standard for publication metadata. Though divert in nature, the overall structure of the two present many shared common characteristics. The Crosswalk also effectively maps the dissimilarities in structures from one to the other. However, examination from a micro-aspect, the two guidelines’ fundamentally different purposes of data use result in differences of information selection and creation. While the current mapping guideline is unable to provide clear instruction for these differences, ONIX-converted MARC records will continue requiring professional reviews and correction.

**Conclusion**

The increased popularity and use of eBooks call for a structured strategy for information preservation of this newly emerging format of digital materials. To understand current achievement and help identify possible recommendations, this paper studies the conversion of ONIX for Books records for MARC metadata, as one is the most popular information management systems among the book publishing industry and the other is the most widely used information metadata system among libraries. Analysis reveals that though current mapping guideline resolves majority of the differences between the two systems, the remaining ambiguity might still be possible cause of critical
errors in the resulting MARC records. Inconsistency of record quality thus is proven as a likely outcome of the conversion. Though the mapping technique for ONIX for Books in support of MARC records already achieves a high success, it still needs continue improvement for further clarification on the “gray areas.” Until then, ONIX for Books is just a shortcut that helps publishers, vendors and libraries save times on metadata creation, rather than an effective contribution towards eBooks’ preservation effort.
References


Library of Congress. (2009). “What is a MARC record, and why is it important?” Retrieved from http://www.loc.gov/marc/umb/um01to06.html#part1


Retrieved from:


OCLC. OCLC Metadata services for publishers. Retrieved from:
https://www.oclc.org/content/dam/oclc/partnerships/material/metadata/213918usb_services_for_publishers.pdf

Preserving digital information: Report of the task force on archiving of digital information. Retrieving from


