Mapping ONIX to MARC

Carol Jean Godby

Research Scientist
OCLC

A publication of OCLC Research

Mapping ONIX to MARC
Carol Jean Godby, for OCLC Research

April 2010

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 1-55653-380-2 (978-1-55653-380-8)
OCLC (WorldCat): 588972377

Please direct correspondence to:
Carol Jean Godby
Research Scientist
godby@oclc.org

Suggested citation:
Godby, Carol Jean. 2010. Mapping ONIX to MARC. Report and crosswalk produced by OCLC Research. Available online at http://www.oclc.org/research/publications/library/2010/2010-14.pdf (report) and http://www.oclc.org/research/publications/library/2010/2010-14a.xls (crosswalk).

Contents

Executive Summary 6

Acknowledgement 7

Introduction 8

1.0. Metadata for Libraries and Publishers 8

1.1. ONIX 2.1 and MARC 21 8

1.2. Mapping ONIX to MARC and back again 12

1.3. A crosswalk from ONIX to MARC 14

1.4. OCLC's implementation 14

2.0. A Closer Look at Some of the Maps 16

2.1. Identifiers 16

2.2. Supplementary text 17

2.3. Subjects 19

2.4. Taking stock 21

2.5. Some challenges 22

3.0. Some Lessons 27

References 31

Tables

The following tables are available in the *ONIX–MARC Crosswalk* spreadsheet available online at http://www.oclc.org/research/publications/library/2010/2010-14a.xls.

Table 1. ONIX

Table 2. ProductForm

Table 3. ProductContent

Table 4. ContributorRole

Table 5. Edition Type

Table 6. Extent

Table 7. BISAC SH

Table 8. Audience Code

Table 9. Audience Range

Table 10. Description

Table 11. Measure

Table 12. Country Codes

Table 13. State Provinces

Figures

Figure 1.1. Correspondences between ONIX and MARC records 10

Figure 2.1. The EAN represented in ONIX and MARC 17

Figure 2.2. Supplementary text represented in ONIX and MARC 18

Figure 2.3. A subject heading in MARC and ONIX 19

Figure 2.4. Maps involving BISAC codes and subject headings 21

Figure 2.5 a and b. ONIX and MARC *Contributor* relationships 23

Figure 2.6. Titles in ONIX and MARC. 25

Figure 2.7. Physical descriptions in ONIX and MARC 26

Executive Summary

This document presents an interpretation of a crosswalk from ONIX 2.1 to MARC 21 developed by OCLC and made publicly available from the OCLC Web site and EDItEUR[1]. This work represents a major upgrade in the statement of how data for bibliographic description can be exchanged between two standards that are widely used in the library and publishing communities. The discussion considers practical outcomes and identifies theoretical and conceptual issues that will inform the next major revision of this strategically important relationship.

[1] EDItEUR. http://www.editeur.org/.

Acknowledgement

The crosswalk described in this article was developed by my OCLC colleagues Renee Register and Bob Pearson and implemented by my project team. I am solely responsible for any errors of interpretation in this discussion.

Introduction

1.0. Metadata for Libraries and Publishers

ONIX, or Online Information Exchange, is a family of international standards for communication between computer systems, developed and maintained by EDItEUR. ONIX standards cover a range of applications related to product description for publications of all types and in all media; intellectual property rights in media content; usage permissions and prohibitions; and the registration of standard identifiers for works and their manifestations. *ONIX for Books* was the first and remains the most widely adopted. It is supported by the New-York-based Book Industry Study Group (BISG) and the UK's Book Industry Communication (BIC), as well as by a growing number of implementation groups in other countries. *ONIX for Books* is designed to promote electronic data interchange between publishers and booksellers. A typical record includes a bibliographic description with elements such as titles, authors, subjects, publishers, and unique identifiers. Other elements describe sales and distribution rights, pricing information, and availability.

Though ONIX data is used primarily to facilitate the movement of materials through a supply chain, it is visible to customers when they read descriptions and reviews on the Web site of an online bookseller or when they visit a bookstore and browse displays of books and DVDs arranged by subject. A visitor to a library would reap many of the same benefits from MARC records, or Machine Readable Cataloging records (LC 1999), making use of a forty-one-year-old standard, variants of which have been adopted by the library community in many countries, initially to generate electronic versions of catalog cards, and today as the basis for online public access catalogs.

1.1. ONIX 2.1 and MARC 21

Corresponding ONIX and MARC records are shown in Figure 1.1 below for a mystery book published in 1988. The ONIX record shows only elements used for bibliographic description because this is the point of overlap with the MARC record. Missing are ONIX elements that

record territorial rights, sales restrictions, availability, and other information pertaining to e-commerce.

The diagram employs the convention that will be used in Section 2 to visualize the detailed results of the mappings from ONIX to MARC.

- The red text represents the unique data values that can be copied from ONIX to MARC with, at most, only slight changes. For example, the <RecordReference> and <ProductIdentifier> elements are copied verbatim to corresponding MARC fields, while the <TitleText> is formatted with different capitalization rules when it is copied to the MARC record.
- The blue text indicates purely structural differences between the two records. The structure of the ONIX record is enforced by an XML schema for ONIX 2.1. MARC records can be realized in more than one form, but Figure 1.1 is a human-readable, tab-delimited version of the ISO-2709 syntax. Reading from left to right, there are three kinds of *signposts* (LC 2009) in a MARC record: a *field*, named by a *tag*, which, except for the Leader, is a three-digit number ranging from 001 to 999; a set of two *indicators*, which contain coded instructions for interpreting or manipulating the data associated with a field; and a set of one or more *subfields* such as $a, which contain data values.
- The green text represents coded data that is not copied but mapped. Sometimes the map is simply a one-to-one correspondence from the ONIX to the MARC code. For example, the ISBN-13 is represented in ONIX as a <ProductIdentifier> with an <IDValue> of 3 and in MARC as an 024 field whose first indicator value is also 3. But other relationships are not so straightforward. For example, the ONIX <ProductForm> value of BB tags this record as a description of a printed book, but the same information is conveyed in the corresponding MARC record as a set of values in the Leader, 008, and 300 fields. The details of these relationships are discussed in Section 2 of this article, but here I simply want to demonstrate that the sections of ONIX and MARC records devoted to bibliographic description are broadly similar and should be shareable across applications in the library and publishing communities.

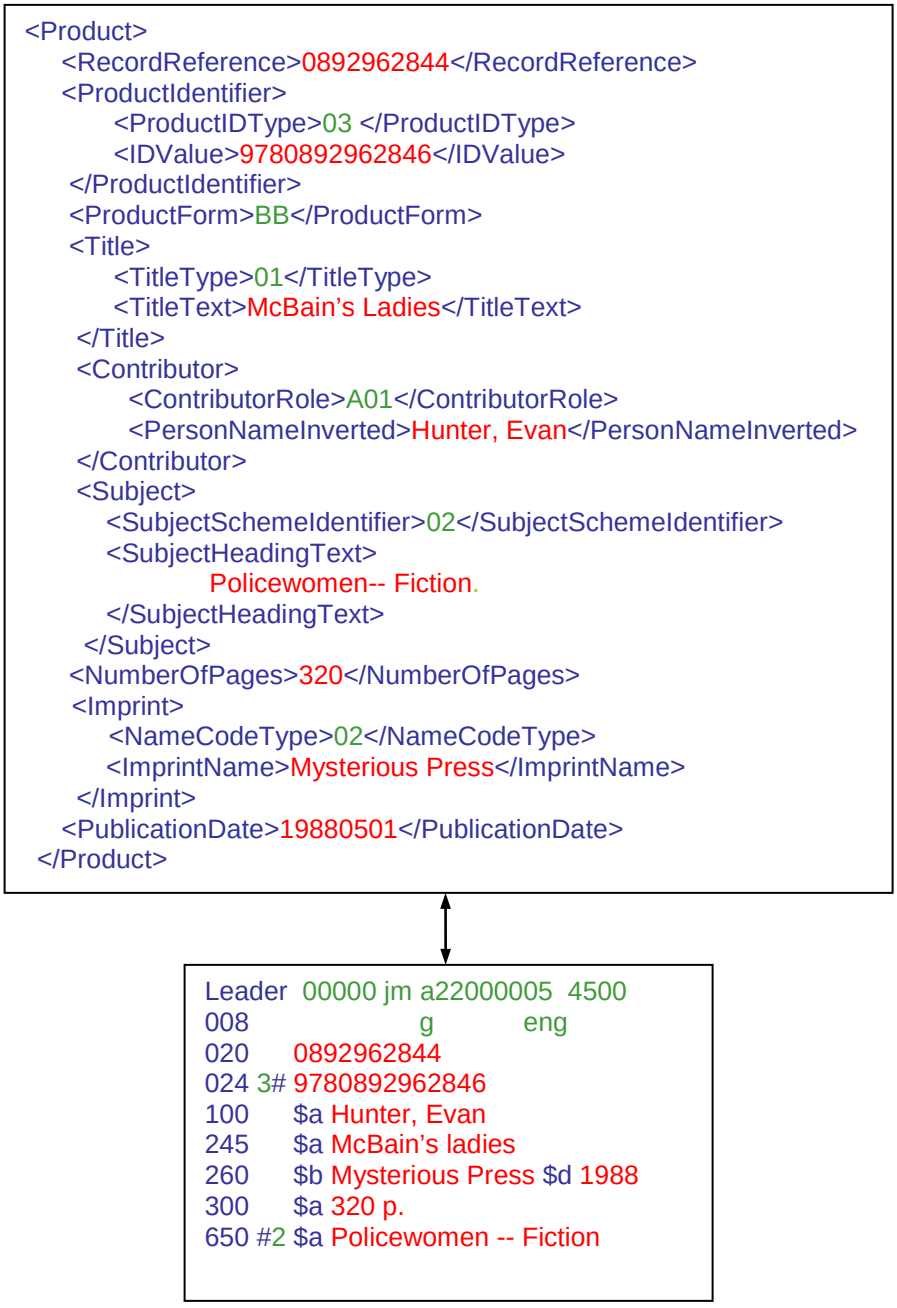

Figure 1.1. Correspondences between ONIX and MARC records

ONIX and MARC have obvious similarities because both standards ultimately serve to connect published works with their audience. But they are not the same standard, nor is one derived from the other. The two standards are structurally and semantically different because they support different needs and communities of practice. ONIX is an international standard for transmitting data about published items that makes no assumptions about how the data will

be used, enhanced, or manipulated by the receiver. MARC is a family of closely related standards that are used internationally to support library applications and exchange information in the library community.

But because the MARC record is intended to be consumed as a package that must meet the expectations of particular applications developed in the library community, it has additional content and formatting beyond what is required for a more agnostic transmission of data. First, the MARC record is formatted according to punctuation rules defined by the International Standard Bibliographic Description, or ISBD (IFLA 2007). In Figure 1.1, ISBD rules specify that only the first word in the 245 (title) field is capitalized and that the abbreviation 'p.' is appended to the number of pages cited in the 300 $a field. Second, the MARC record conforms to cataloging rules that are defined for particular linguistic or geographic regions. In the United States and Canada, MARC records typically obey the *Anglo-American Cataloging Rules, Second Edition*, or *AACR2* (JSC-AACR 1988), which ensure that the record is appropriate for use in online library catalogs. Special attention is paid to the construction of *access points* that are designed to enable the user of such catalogs to retrieve a bibliographic record from a search query. In the MARC record shown in Figure 1.1, the access points are the 100 field, which lists the primary contributor; the 245 field, or title; and the 650 field, or subject. But the data in this example is too sparse to show that these fields are more complex than alternative expressions of the concepts in the ONIX <Contributor>, <Title>, and <Subject> composites. Because of their special function in library catalogs, access points may contain data mapped from multiple ONIX composites, such as physical medium descriptors in the 245 field or titles in the 100 field. The details are described more fully in the next section.

To summarize, the issue of mapping between ONIX and MARC appears at first glance to be straightforward. It is easy enough examine the data shown in red in Figure 1.1 and discover that the ONIX <Subject> element corresponds to the MARC 650 field; the ONIX <Contributor> element matches the MARC 245 field; and so on. But it is important to be clear about terminology and levels of abstraction. And when we are, the problem space becomes much more complex. First, the terms *ONIX* and *MARC* are informal designations for standards that may have multiple syntactic realizations and versions. In this article, *ONIX* is a cover term for ONIX for Books 2.1 and *MARC* represents the ISO-2709 representation of MARC 21. Another issue is that the MARC target, but not the ONIX source, may have special formatting, such as ISBD punctuation, which must be introduced somewhere in the conversion process. Thus the technical problem is to design data models and conversion software that preserve the essential relationships between the two standards across different versions, syntaxes, and formatting conventions, or permit easy revision when fundamentally new changes are introduced.

But if the goal is to ingest records into databases that support library functions, one more step is necessary. The semantic correctness of the record must be verified by checking for the presence of required data elements. Unfortunately, semantic verification has no universal standard. Following established practice, OCLC's semantic checks for MARC records are based on AACR2, which has an Anglophile perspective. But OCLC is also monitoring progress on the proposed replacement for AACR2, Resource Description and Access (RDA), which promises a more international scope and permits greater interoperability with non-MARC standards (RDA 2010). In addition, we perform a rudimentary semantic check on the ONIX records based on the recommendations of the New York-based Book Industry Study Group (BISG) and have plans to incorporate those of UK-based Book Industry Communication (BIC). Nevertheless, semantic validation is the most unsettled component of our translation model, and we welcome advice from experts in the standards communities for improving it.

Though the two standards are structurally and semantically different, the relationship between ONIX and MARC is strategically important. For one thing, libraries constitute about 12% of the publishers' market because they purchase many items from booksellers. Once the items are acquired, the accompanying descriptions created by the publishing community must be integrated into library databases.

1.2. Mapping ONIX to MARC and back again

To accomplish the goal of merging publisher and library metadata, ONIX data must be mapped to MARC. To be done efficiently, this work should be automated, which implies that the goal of electronic data interchange articulated by the proponents of ONIX would ideally extend beyond the vendor and publisher communities to include a community with a different standard. But the data exchange between publishers and libraries need not be unidirectional. Records originating from publishers and booksellers contain authoritative information about titles, authors, publisher names, and unique identifiers such as ISBNs, as well as links to reviews, summaries, tables of contents and sample chapters that create a context for customers who access bookseller websites to find related works. But library records create valuable context of their own, such as classifications, more detailed subject headings, and links from the names of authors to resources that establish their real-world identities. Library records may also reflect models of bibliographic description that formally associate different manifestations of the same intellectual work, such as older and more recent editions, or hardback and paperback copies of the same book (IFLA 2010). Thus if data from libraries and publishers could be integrated, essential information could be transferred easily between the two formats and the result might be an enhanced record that both communities would consider more valuable. As a result, it is urgent to identify the common ground in the two standards and devise software solutions that merge ONIX and MARC records automatically.

Some of these ideas are being explored in OCLC's *Metadata Services for Publishers* (OCLC 2009) project, which was released in mid-2009 and continues to evolve. The goal is to produce high-quality records for publishers and libraries while addressing inefficiencies in the management of metadata in the publisher supply chain. A fully automated process obtains ONIX records from U.S. publishers and vendors, translates them to MARC21 (the version of MARC which is now adopted in a number of major countries, particularly in the English-speaking world), and enhances them with data mined from similar records in OCLC's WorldCat® database of tens of millions of bibliographic records, the largest such collection in the world (OCLC 2010b). The resulting MARC record is made available to libraries; and in a separate stream, the same record is translated back to ONIX for delivery to the publishing community.

This project requires a robust *crosswalk*, or a set of semantic correspondences, from ONIX to MARC21. Such crosswalks have been proposed before, but the *Metadata Services for Publishers* project team determined that they were incomplete, inconsistent, or out of date. One of the early deliverables was thus a major overhaul of the ONIX-to-MARC crosswalk and the corresponding record-builder algorithm available from the Library of Congress (LC 2000), the most comprehensive version now publicly available. The updated crosswalk is currently defined for ONIX for Books 2.1 (EDItEUR 2009a), but updates for ONIX 3.0 (EDItEUR 2009b) are underway. The crosswalk has been in use at OCLC for the past six months in the *Metadata Services for Publishers* and other metadata management projects. We are now making the crosswalk publicly accessible from EDItEUR so it can be vetted by the library and publishing communities and, hopefully, improved.

Our goal was to define the intelligence for a commercial-grade system that can map the bibliographic descriptions contained in ONIX and MARC records and satisfy three requirements.

First, MARC records produced by the system pass MARC and AACR2 validation rules. Second, ONIX records produced by the system must be syntactically valid according to the XML schema published by EDItEUR and pass a validation check based on the Book Industry Study Group best practices guidelines (BISG 2005), which require the presence of the core elements in a bibliographic description, such as the elements that denote identifiers, titles, creators, publishers, subjects, and physical dimensions. Finally, ONIX and MARC records produced by the system must, wherever possible, survive a roundtrip translation, which maps from the source to the target and back to the source. A robust roundtrip translation serves a legitimate business need in the *OCLC Metadata Services for Publishers* project because ONIX records obtained from the publisher supply chain must be returned in ONIX to their suppliers, usually enhanced with data fields mapped from MARC records. But the roundtrip also tests the

integrity of the conversion process because it prohibits data values from being dropped or changed.

1.3. A crosswalk from ONIX to MARC

The crosswalk is represented as a human-readable Microsoft Excel spreadsheet that can guide an implementation of software that translates from ONIX to MARC. Organized into fifteen worksheets, the most important worksheet is labeled *ONIX* (also referred to in this paper as Table 1). The subordinate worksheets contain details required to fully implement some of the maps. They typically specify how ONIX codes may trigger maps to multiple MARC fields or spell out complex conditional logic that cannot be stated succinctly.

Since the crosswalk is designed to be freestanding, it should be possible to interpret the instructions largely without consulting reference works such as the Library of Congress' *MARC 21 Format for Bibliographic Data* (LC 1999) or EDItEUR's documentation for ONIX 2.1 (EDItEUR 2009a). In Table 1, the focus of most of the discussion in this article, the first three columns give a schematic representation of the ONIX element or hierarchically structured composite to be mapped, which contain one element per row and are identified using the human-readable ONIX reference element names. The data in the fourth column is a brief description of the element's semantics. The next column specifies whether the ONIX element is mandatory or optional according to the Book Industry Study Group's recommendations (BISG 2005) for freestanding elements, or the ONIX 2.1 schema (EDItEUR 2009a) for elements that appear in an ONIX composite.

The last column contains the corresponding MARC element, along with a set of human-readable instructions for coding a map from the ONIX source. Some of the instructions are simple, such as the one-to-one transfer of codes required to map from the ONIX <NotificationType> element to the Encoding Level subfield of the MARC Leader field. Others describe relatively transparent Boolean logic. For example, the map from the ONIX <LanguageCode> targets either the *Default Language of Text* subfield of the MARC 008 field, or the 041 field, depending on whether the language is primary or secondary. But sometimes the instructions are extraordinarily complex, semantically opaque, and require reference to secondary worksheets. For example, the ONIX <ProductForm> element triggers a map to as many as six MARC fields, some containing up to fifteen subfields.

1.4. OCLC's implementation

The ONIX-to-MARC crosswalk is fully implemented in OCLC's Crosswalk Web Service. The technical details have been described elsewhere (Godby, et al. 2008a and 2008b), but here it

is instructive to mention one feature of the design that makes it easy to define the translations and manage the variability that will occur as different versions of ONIX are reconciled with multiple versions and syntactic representations of MARC. The fundamental concept in the translation model is a *map*, which roughly corresponds to a row in a crosswalk defined by a metadata standards expert. A set of related maps is referred to informally as a *mapping*. For example, the mapping for identifiers contains a map for the 10-digit ISBN, whose ONIX source is <ProductIdentifier> (with a <ProductIDType> value of 01) and whose MARC target is MARC 020 $a. As in the human-readable spreadsheet representation, the map has a source, a target, and, optionally, some conditional logic. But the map is otherwise self-contained because no other maps refer to it. This feature makes it easy to discuss individual maps, as I do in the next section.

More substantively, the self-contained property of maps implies that a crosswalk, as well as the machine-processable translation derived from it, is built up dynamically from an unordered set of maps that can be large or small, depending on the task at hand. For example, a translation that implements all of the maps defined in the crosswalk might be used to translate comprehensive ONIX records to archival-quality MARC records and vice-versa, both of which would be candidates for submission to a semantic validation process. But since both kinds of records originate in data streams to which changes may be applied, the atomic map can also be interpreted as a logically transparent carrier of incremental change.

For example, in the *Metadata for Publishers* process flow, subject headings mined from MARC records are applied to matching ONIX records. And tables of contents or other supplementary data obtained from ONIX records are applied to corresponding MARC records. In these cases, only the maps from the ONIX <Subject> and <OtherText> composites are involved, not complete records. The map concept also provides a clear conceptual framework for implementing the changes required for ONIX 3.0, despite the fact that the new version is not backwardly compatible with ONIX 2.1. Fortunately, the crosswalk does not require a complete rewrite because maps for new elements are added, maps for deprecated elements are deleted, and many maps for elements that retain their meaning but are repositioned in the ONIX record structure can be easily modified.

In the next section, I describe some of the most important maps required for bibliographic description in ONIX and MARC. The discussion is not comprehensive and it is not intended as a 'record-builder' algorithm. Instead, my goal is to provide an interpretation that attempts to shed light on why the maps take the form that they do, starting with simple examples and progressing to the most complex and problematic ones. This organization reveals issues that will need to be resolved by future work, perhaps involving MARC and ONIX standards committees.

The OCLC project team started with the assumption that ONIX could be mapped to MARC, given a sufficiently rich definition of *map* that permits many-to-one relationships or conditional logic. Once the equivalences between source and target elements are established, the associated data found in a record is copied, manipulated algorithmically, generated as constant data, or mapped from a corresponding table. These operations are rich enough to create records that can be submitted to syntactic or semantic validation in a commercial-grade implementation. But since these operations essentially establish syntactic and lexical equivalences, they are less effective where ONIX and MARC have major semantic differences.

2.0. A Closer Look at Some of the Maps

Mappings between ONIX and MARC are possible because the two standards are compatible and, in some respects, complementary. But this is not an accident. Since the ONIX standard was proposed some thirty years after MARC was first adopted by libraries, it was informed by lessons learned from users of MARC. Members of the library community participate in the Book Industry Study Group Metadata Committee, an advisory body based in New York that recommends best practices for the deployment of the ONIX standard in North America. Among the recommendations is a proposal to certify data providers whose records contain a list of required ONIX elements. Most of these elements are central to the bibliographic descriptions created by members of the library community; they include unique identifiers, titles, creators, subjects, material types and physical descriptions, and language of the content. Others are tailored to important events that happen to a book or DVD as it moves through the bookseller supply chain, such as the ONIX elements indicating the return policy, territorial rights, the number of items that fit in a pack, and the name of the supplier. The list of elements and the BISG recommendations are summarized in *Product Metadata Best Practices for Senders (*BISG 2005). In this section, I will examine how ONIX elements that form the core of the bibliographic description fare in the mappings to MARC.

2.1. Identifiers

Figure 2.1 shows fragments of ONIX and MARC records illustrating the result of applying the map defined in the crosswalk for the EAN-13 article number, which is used by retail stores for inventory control and checkout. The essential information is given in Row 17, Column E of Table 1, *ONIX,* in the ONIX-to-MARC crosswalk (OCLC 2010a). This, plus the preceding two rows establish a symmetric relationship between the ONIX <ProductIdentifier> composite and the MARC 024 field.

The MARC record fragment is shown in the box at the top of the diagram and the corresponding ONIX data is below it. The color coding means the same as it does in Figure 1.1.

Structural differences are shown in blue; data values that are copied are shown in red; and data values that are mapped are shown in green. In this example, the ONIX <ProductIdentifier>, which has the required children <ProductIDType> and <IDValue>, maps to the MARC 024 field. The value '03' of <ProductIDType>, shown in green, is mapped to '3' in the MARC second indicator. And the data contained in the ONIX <IDValue> element, shown in red, is copied to the MARC $a field.

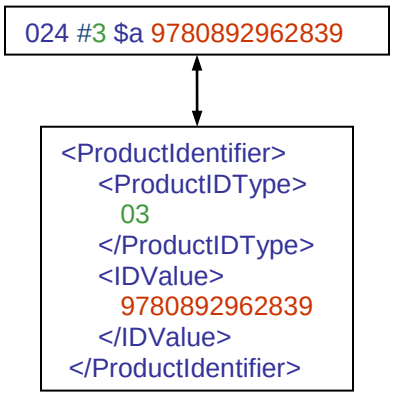

Figure 2.1. The EAN represented in ONIX and MARC

The <ProductIdentifier> is a mandatory element according to the BISG Best Practices Guidelines. The value of the ONIX <ProductIDType> is a code that specifies an identifier such as EAN, UPC, OCLC number, ISBN, or nine other values; the corresponding information is represented in MARC in the 020 or 024 fields and a range of values in the second indicator. Given that this is critical information in the library and publishing communities, it is fortunate that the mapping preserves most of it. The one exception is the difference between ten-digit and thirteen-digit ISBNs. The ONIX representation distinguishes them with different values for <ProductIDType> and for trade purposes it is critical to do so, particularly during the long period during which currently traded products have had to carry both forms. But the two types of ISBN have the same representation in $a of the MARC 020 field. To distinguish between them, OCLC's translation software must examine the data and count the number of digits, a workable if inelegant solution.

2.2. Supplementary text

Figure 2.2 shows record fragments produced from the instructions for mapping supplementary text given in Table 10, *Description*. These instructions map ONIX elements containing tables of contents, biographical notes, reviews, flap comments, back cover copy, and summaries. In this example, the ONIX <OtherText> element, with the required sub-elements

<TextTypeCode> and <Text>, is mapped to MARC 520; the value of the second indicator, shown in green, identifies the type of material as a summary; and the data in the <Text> element, shown in red, is the copied text.

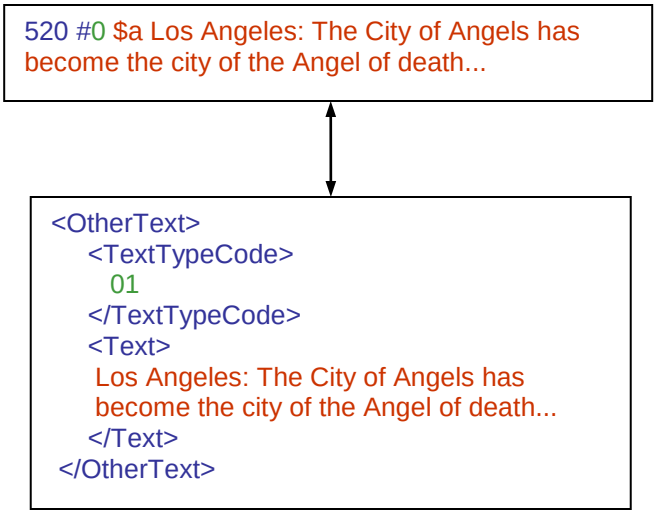

Figure 2.2. Supplementary text represented in ONIX and MARC

This material is an important source of enrichment from ONIX records, which may map to the MARC 500, 520, or 545 fields. Because of the diverse MARC targets, the map is similar in complexity to the maps involving the <ProductIdentifier> element, except that there is no need to refer to the text. As a result, the map may be slightly more straightforward, but no assumptions can be made about how the text is represented. For example, tables of contents may have hard-coded indentations for chapter and section headings. The text may be explicitly enclosed in XML <CDATA> elements, indicating with formal markup that the text should be displayed as formatted. Or it may be a simple ASCII stream. If the text contains non-Roman characters, the character encoding may be difficult for software processes to identify unambiguously.

Another problem with the supplementary material is that the ONIX standard makes a distinction between resources that correspond to readable text and those that are represented in non-text media, such as images or audio files, a distinction that the ONIX designers now consider to be arbitrary. References to non-text media are encoded in the <MediaFile> element, which forces creators of ONIX records to make difficult decisions about recorded speech and turns the relationship to MARC into a complex many-to-many mapping. Perhaps because of this problem, *ONIX for Books 3.0* retires the <OtherText> and <MediaFile>

composites and draws the distinction slightly differently. Written text, including summaries, cover blurbs, and review quotes, is transmitted as part of the record in the <TextContent> composite, while third-party content in any medium, such as a magazine article or a radio broadcast, is linked from the <CitedContent> composite. Of course, the revamped relationship to MARC will have to be expressed as an even more complex set of many-to-many maps, since the MARC standard cannot change in tandem to accommodate this significant redesign of a key set of ONIX elements.

2.3. Subjects

The record fragments shown Figure 2.3 illustrate one outcome of the mapping between ONIX and MARC subject elements, another relatively straightforward relationship. The relevant maps are defined in rows 107-115 in Table 1, *ONIX*.

In this example, the code 04 in the <SubjectSchemeIdentifier> field identifies the data in the <SubjectHeadingText> as a Library of Congress subject heading. The full list of identifier codes refer to the subject heading schemes commonly used in the library community, such as National Library of Medicine, Sears, and Dewey, as well as those used by publishers and booksellers.

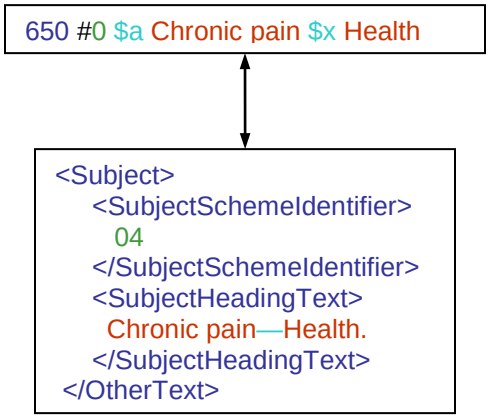

Figure 2.3. A subject heading in MARC and ONIX

The alternation shown in turquoise is unique to this example and is intended to show that the two MARC subfields are collapsed to a single data value and separated with dashes when they are mapped to the ONIX <SubjectHeadingText> element, observing an ISBD punctuation rule for subject headings. Other LCSH subfields can participate in this alternation, as the crosswalk shows. The result is a text string formatted by LCSH conventions, but the subfield sources of the data is lost if the <SubjectHeadingText> is mapped back to MARC. In other

words, a human reader or a machine process cannot determine if *Health* came from $x or some other MARC subfield.

The data flow for the OCLC Metadata for Publishers project also requires maps to MARC from BISAC headings, which are maintained by the Book Industry Study Group. The need to handle BISAC introduces a few semantic complexities to the ONIX to MARC relationship, as defined in row 116 of the crosswalk, all of which can be translated from ONIX to MARC and back again without loss of information. Publishers make a distinction between the *BISAC code*, used in the supply chain to facilitate electronic data exchange; and the corresponding *BISAC text*, which is used to populate documents intended for human consumption and to construct store displays arranged by subject. This distinction is retained when the ONIX data is mapped to MARC, but it acquires a different meaning. As noted above, the BISAC text corresponds to a subject heading when it is mapped to a MARC record. But the code is interpreted as a classification when it is the primary subject assignment recorded in the ONIX <BasicMainSubject> element.

Though subject headings and classification codes are logically indistinguishable in the publisher supply chain, librarians view them as two separate but related tools that organize the world's recorded information for the purpose of making it discoverable and retrievable (Chan 2007, 3). According to OCLC's classification experts, BISAC can be interpreted as a classification scheme with two levels of hierarchy. The first three letters of the BISAC code identify one of fifty-one top-level subject categories or genres (BISG 2009), such as *History*, *Poetry*, *True Crime*, *Study Aids*, or *Bibles*. The six-digit code that follows the prefix identifies a narrower subject, such as *Poetry/African* or *Poetry/Inspirational & Religious*. To make BISAC codes more consistent with longstanding library classification schemes, the three-letter BISAC prefix is mapped to 072 $a, the MARC field containing a subject category code, while the extension is mapped to $x.

These details are illustrated by the MARC fragment on the left in Figure 2.4. BISAC subject codes listed in ONIX <Subject> composites are converted to their corresponding textual values and represented in the output MARC record as a list of 650 fields, as described above. The MARC fragment on the right in Figure 2.4 is essentially identical to that shown in Figure 2.3, except for the indicator value of 07 that identifies the BISAC source of the heading. Note the constant data applied to the $2 subfield shown in the MARC fragment on the left, which also annotates the BISAC source of the data values.

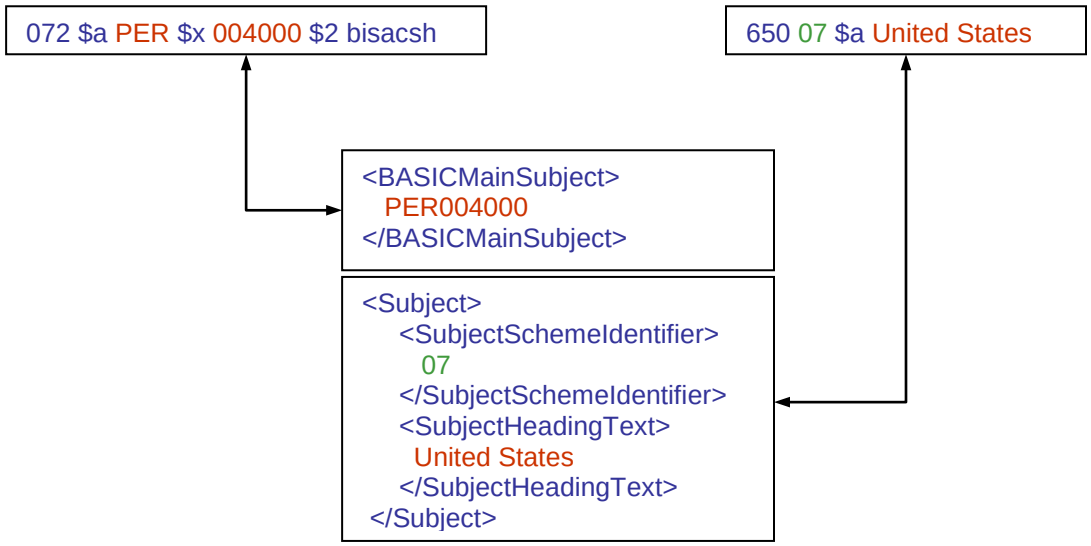

Figure 2.4. Maps involving BISAC codes and subject headings

The BISAC code can be parsed for additional information that is used to populate MARC 008 fields such as *type of record*, *audience*, or *literary form*. Details are spelled out in Table 7 (*BISAC SH)*.

2.4. Taking stock

This is good place to make some high-level observations about the ONIX to MARC relationship before discussing some more problematic maps. All of the maps described so far involve required elements according to the BISG recommendation and the affected elements survive the translation to MARC well enough to support commercial-quality processes. This is perhaps a consequence of the fact that the semantics of the elements are essentially preserved. For example, the unique identifiers—the ISBN, the ISSN, the OCLC number, and so on—mean exactly the same thing in ONIX and MARC. So do the tables of contents, author biographies, reviews and other data transmitted through the <OtherText> element, despite the fact that this data sometimes poses problems with display and other machine-to-machine processes.

But one cause for concern is that even in the simplest examples, it is necessary to refer to and possibly change the data, leading to loss of information if the translation has to survive a round trip. For example, when the affected elements involve subject headings, the data must be stripped of ISBD punctuation conventions when the source is MARC and more than one subfield is present. But the information about how to restore this change is lost if the

resulting ONIX data is later mapped back to MARC, a problem that highlights subtle differences in granularity even among elements that been carefully studied by experts tasked with aligning the two standards. This is a pervasive issue with the ONIX-MARC relationship, about which I'll say more in last section of this article.

2.5. Some challenges

Three of the most important but most complex mappings in the ONIX to MARC relationship involve the ONIX <Contributor> element and <Title> elements, as well as the ONIX elements containing details about physical characteristics, such as <ProductForm> and <Measure>. The complexity can be traced to complex syntactic relationships, major semantic differences, and legacy elements appearing in both standards.

2.5.1. Authors and contributors

Figure 2.5 shows the results of two maps that underpin the relationship between the ONIX and MARC authorship elements, which are defined in rows 53-75 of Table 1, *ONIX.* The maps must resolve two issues. First, AACR rules makes a fundamental distinction between the person chiefly responsible for creation of the intellectual or artistic content of a work, which is described in the 100 field; and a secondary contributor, described in the 700 field. This distinction does not exist in ONIX, but can be approximated by the values in Codelist 17, which specify a set of <ContributorRole> codes such as *Screenplay by*, *Lyrics by*, *Illustrated by*, and *Introduction by*. The value A01, or *Author of a textual work*, is designated as the primary creator in a MARC record and is mapped to the 100 field; <Contributor> composites containing other values are mapped to 700 fields. If an ONIX record contains more than one <Contributor> with a <ContributorRole> of A01, the first is mapped to 100 and the others are mapped to 700. The effects of these rules are visible in the two record fragments shown in Figure 2.5. Abraham Smith is a primary author of a work of text and is described in a MARC 100 field, while the director Peter H. Hunt is described in a 700 field.

The second issue is the syntax or format of the name. The required $a subfield in the MARC 100 and 700 fields must contain the surname, but more typically it has a string containing a surname and a forename, represented in inverted order. Optional MARC subfields contain fuller versions of the name mentioned in $a, family or dynasty names, titles, and affiliations. All of this information is available to be mapped from the ONIX <Contributor> composite. OCLC's implementation of the ONIX to MARC crosswalk checks for the <PersonNameInverted> element in an incoming record. If it is not present, an inverted name is constructed from the individually tagged name components. The two name formats are illustrated in Figure 2.5 a

and b. The resulting name string in the MARC record is essentially the same regardless of the ONIX source.

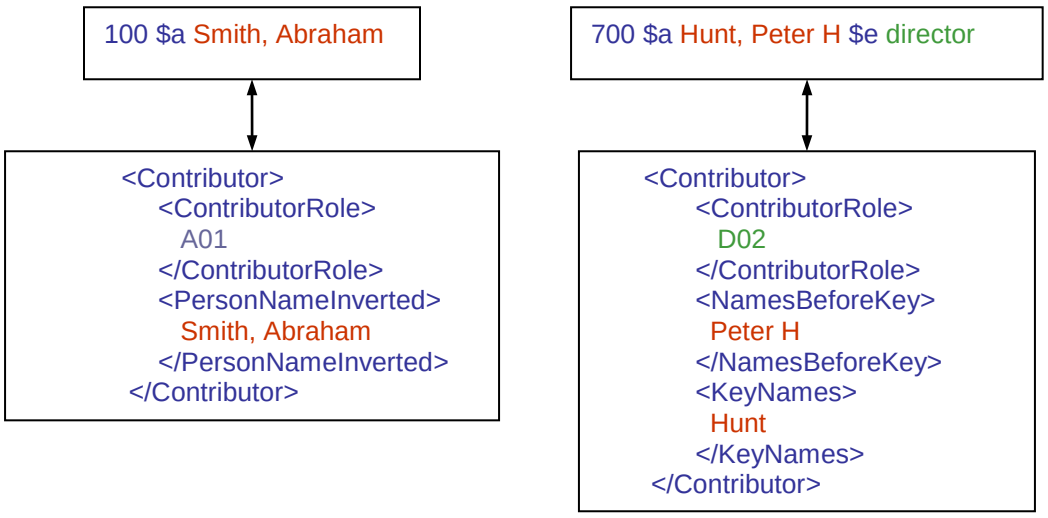

Figure 2.5 a and b. ONIX and MARC *Contributor* relationships

This logic accounts for much of the detail in the ONIX and MARC Contributor elements. But to understand why it is so complex, it is instructive to look beyond superficial structural differences. Part of the complexity can be ascribed to the fact that <PersonNameInverted> is an exception to the principle that ONIX is usually agnostic about the punctuation and formatting of data values. In this case, the data value does have a required format, which is conveniently annotated by the name of the element.

More significantly, the complexity can be traced to semantic differences in the definitions of the *Contributor* elements in the two standards. In ONIX, the <Contributor> element has primary importance. Accordingly, it is self-contained, freestanding, and the child of the topmost node, the <Product> element. The BISG Best Practices guide states emphatically that the <Contributor> is of paramount importance to booksellers because

> …the author of a book is often the most recognizable "brand" of our products that consumers know. In some subject categories, other key data points such as title, publisher, series, etc. are almost irrelevant when compared to the importance of the name(s) of the contributor(s) to that product. The title of a new novel by John Grisham, for example, is not the piece of data that will sell that book. (BISG 2005, 23)

But in a MARC record, the *Contributor* data is subordinated to the meta-concept of *access point,* i.e., the name, term, or code under which a bibliographic record may be searched in a computer system. The main access point for a MARC record is the 100 field, which contains the most pertinent data about a primary author and the works he or she has created, such as a standard form of the name, birth and death dates, the title of a work, and its publication date. Essentially the same information for secondary contributors, such as co-authors, translators, or illustrators, is contained in the 700 field, or the *added entry*. Since access points have no exact correlate in the ONIX record, the concept is lost completely when a MARC record is mapped to ONIX. In addition, the data itself is problematic because the MARC subfields of these fields that contain descriptions of contributors and titles are less granular than the ONIX counterparts. Moreover, these problems are duplicated in the main and added entries that feature corporate authors or uniform titles.

Still more complexity in the maps involving *Contributor* elements can be attributed to the fact that ONIX and MARC records both carry archaisms that testify to their constant evolution. For example, an ONIX biographical description element can occur either in the <OtherText> or <Contributor> composites, so the crosswalk has to account for both possibilities. Similarly, an annotation about the contributor's role—primary author, secondary author, illustrator, translator, editor, and so on—can occur either in MARC 100 or 700 $e as a code, or in $4 as text. Both standards also contain obsolete elements, which can be safely ignored if records to be translated conform to the latest versions, but are in danger of getting dropped if they do not. And since the ONIX 3.0 revisions do not address these issues, the maps have the potential to grow even more intricate as both standards continue to evolve independently.

2.5.2. Titles

The *title* elements in ONIX and MARC, which are defined in rows 43-48 of Table 1, *ONIX*, present two of the same issues as the *contributor* elements. First, the concept of *title* is primary in an ONIX record, while the corresponding MARC fields are interpreted as access points consisting of heterogeneous data. Note, for example, the 245 field shown in the middle of Figure 2.7, which has a $a subfield containing the title text and a $h subfield containing a medium designator for a sound recording.

And as in the case of authorship elements, the ONIX title elements are a combination of unformatted and formatted text fields. For example, the <TitleText> element transmits the title with no assumed punctuation or formatting, while <TitleWithoutPrefix> specifies the alphabetized form of the title. Both versions of the title are mapped. Figure 2.6 shows ONIX and MARC fragments containing a distinctive title and the corresponding translated and alternative titles that have been generated from a description of a Hermann Hesse novel that

is submitted to the crosswalk. In this illustration, the ONIX source for the primary title recorded in MARC 245 $a is constructed from the ONIX fields with special formatting. The other MARC fields are mapped from the ONIX <TitleText> field. These examples show how selections from a flat list of ONIX codes that populate the <TitleType> element identify the strings contained in <TitleText> or the corresponding formatted fields as primary titles (01), translations (06), or alternative titles (05). This information is mapped to MARC fields in the range of 242 to 247, depending on the value of <TitleText>. Only the map to 246 is problematic because ONIX identifies alternative titles, such as *distributor's title*, *title acronym*, *alternative title on front*, or *alternative title on back*, which are lost in translation because MARC does not make these distinctions.

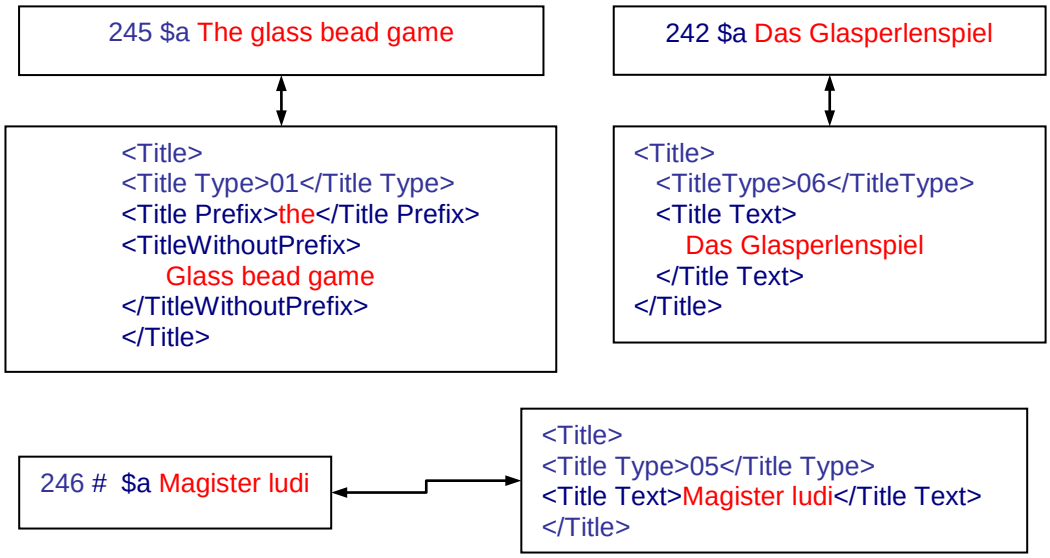

Figure 2.6. Titles in ONIX and MARC.

2.5.3. Physical descriptions

The most complex mapping from ONIX to MARC involves elements that comprise a physical description. Such information is crucial in the publisher supply chain because most of the items bought and sold have physical dimensions that must be weighed, measured, priced, assigned the correct postage, stacked on pallets, or stored in warehouses. More importantly, customers make buying decisions about books based on how big they are, how many pages they have, whether they are illustrated, or whether they are hardback, paperback, or come in a slipcase. Similarly, a request for a DVD is not fulfilled if a Blu-Ray disk is ordered but an older format is received. Librarians have many of the same concerns about objects that must be stored and discovered, but the physical description also serves a long-term archival

purpose. The description must be detailed enough to capture the distinguishing characteristics of multiple editions, or to match machine-readable media with the technology that unleashes its content. And both requirements must be satisfied in perpetuity.

Figure 2.7 illustrates many of these issues through a fragment of matched ONIX and MARC records that describe a compact disc containing selected recordings of arias from Puccini's operas by various artists. The key data element in ONIX is 'AC,' the value of <ProductForm> indicating that the record describes an audio compact disc. The matching values are shown in green, as in the previous diagrams. As Table 2 *(ProductForm)* shows, the <ProductForm> value of 'AC' triggers a one-to many map to six MARC fields—one of which, the MARC 007 field, has twelve subfields. The values in these subfields indicate that the object being described is a mass-produced plastic-and-metal sound disc containing a digitally stored digital recording that can be played at the constant linear velocity of 1.4 meters per second. The 'm' value in the 'Bibliographic level' subfield of the MARC Leader field indicates that the object is an individual standalone item, not a collection or a component part of a series. And the sibling 'j' value in the second position of the Leader marks this object as a sound recording, a designation that is repeated with a textual value in the 245 $h subfield.

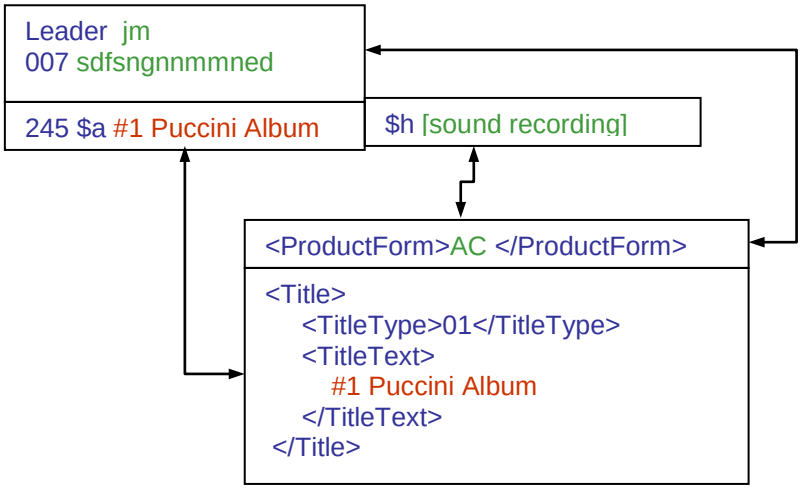

Figure 2.7. Physical descriptions in ONIX and MARC

The topic of user discovery might prompt an attentive reader to wonder why the end-user term 'CD' appears in neither the MARC nor the ONIX description and what consequences this omission will have for buyers and library patrons who are looking a recording of the Puccini arias in this format. Of course, the ONIX term *Audio CD* comes closer than the MARC term *sound recording* or the list of descriptors in the 007 field. Proponents of controlled vocabulary

would argue that the street name *CD* is too imprecise or ephemeral, or that the ONIX term *Audio CD* satisfies no anticipated use because it is neither a controlled nor an end-user term. This is because only the ONIX code 'AC' is controlled; the associated term *Audio CD* is simply a gloss, which is not guaranteed to persist across different languages or use cases. But the description in the MARC 007 field is perhaps overly specified. Though it may accurately describe a particular recording that was manufactured in the United States in 2004, the technical specifications will surely change over time and they could even now be subject to regional differences. If so, this map will have to be replicated many times, with different values for some of the MARC subfields. These problems are due to the fact that much of the key vocabulary in the domain of physical description of published items is not standardized, defined, or decomposed into sub-elements that are useful to all stakeholders.

The physical description elements in a bibliographic record will only get more complex because technology continues to evolve. Indeed, a primary motivation for ONIX 3.0 is to upgrade the description of e-books. An ONIX 3.0 e-book description will have a file size, as it does now, but is this enough information to make an informed buying decision? If some readers want to be assured that a book is thorough and comprehensive, while others might be put off by a book that is too long, would they be better served by a description of the e-book's physical antecedent, which would have a page count and measurements? If so, will the library community accept this information if the curatorial work has not been done to establish the printed book as the true antecedent instead of merely a good match? How should different file formats be described, which is essential information because e-book readers are currently incompatible? And will each format have a separate ISBN even if the content is otherwise identical? Answers to these questions[2] will affect the relationship between ONIX and MARC, and will certainly not simplify it.

3.0. Some Lessons

All of the ONIX data elements involved in the ONIX-to-MARC maps described in the previous section are required elements according to BISG's *Best Practices Guidelines for Data Senders* (BISG 2005). They are among the most important cases that must be handled in a commercial-grade metadata management system that focuses on promoting interoperability among bibliographic descriptions. The resulting crosswalk is commercially viable, but it represents a substantial intellectual effort that will be difficult to maintain, despite an implementation in a software environment that is optimized to manage change.

[2] These issues were discussed in a webcast sponsored by the Book Industry Study Group on October 7, 2009. Slides are available at http://www.bisg.org/events-0-469-bisg-webcastonix-for-books-30supporting-new-metadata-for-ebooks.php.

In this section, I would like to reiterate what makes the maps more complex or brittle than they need to be. To ground the discussion, I will invoke a linguist's perspective and consider, in turn, how the syntax, semantics, and pragmatics of the two standards contribute to the problem but might also suggest a productive way forward.

The mapping problem can't be solved at the syntactic level because the translation must peek at the data.

It is tempting to formulate a translation model for metadata management that starts with a lexical equivalence: for example, ONIX <Subject> is equivalent to MARC 650, or ONIX <ProductIdentifier> is equivalent to MARC 024, and so on. This model implies that translation occurs with perfect synonymy and that a straightforward (or at worst, a complex but manageable) amount of structural manipulation is required to get from the input to the output. In such a model, a mapping from ONIX to MARC would be akin to a structural conversion with a lexical relabeling, which is only one step more complex than that required to produce MARC-XML from the older ISO-2709 MARC syntax.

This simple model is a starting point that can handle much of the routine work in mapping from a metadata source to a closely related target in the ONIX to MARC crosswalk. But as the above discussion has demonstrated, it is difficult to find even a single map that does not require a look at the associated data—which introduces an extra condition in the conversion, a possible loss of data, or a slight shift in meaning.

The semantic differences between the standards should be harmonized.

Though the elements that support bibliographic descriptions in ONIX and MARC have much overlap, the two standards are semantically different. As the discussion of the *Subject* elements in Section 2.3 pointed out, ONIX lacks a distinction between classifications and subject headings. Nevertheless, the ONIX <BASICMainSubject> is a reasonable stand-in for a classification code in a MARC record, while the BISAC text strings can be interpreted as subject heading. Analogously, the ONIX standard lacks the MARC concept of *access point*, but it can usually be reconstructed. As noted, translations involving these elements involve some loss of granularity, but it is not catastrophic even for the ambitious goal of creating records in both standards that pass semantic validation checks.

A more pervasive problem is that the essential synonymy of concepts defined in the two standards is too often obscured. This can happen because a concept such as the physical description of an item in MARC is expressed redundantly and verbosely, reflecting a history of changing cataloging practices that is nearly impossible to map to a newer standard that has

no such need to respect legacy, as the discussion of physical description in Section 2.5.3 demonstrates. More typically, though, slight differences in the representations of the data values present nearly insurmountable obstacles to a determination that the source and target elements in a map are equivalent and, in fact, have the same value.

For example, consider what happens at the stage in the *OCLC Metadata Services for Publishers* workflow that matches an incoming ONIX record to an existing record in the OCLC WorldCat database in an effort to merge the data from the two sources and create an enriched record. The incoming ONIX record has an <OtherText> element coded as a table of contents and the corresponding MARC 505 field has a Table of Contents field containing the same data. But if chapter and section headings are separated by dashes or numbers in the ONIX description, while the MARC description has hard-coded indentations, the two fields will not be merged and the output will have redundant information that only a human reader can easily detect. The same point can be made about author, title, subject, publisher or other fields in the bibliographic description that contain text whose format is not strictly specifed.

Fortunately, many of these data representation issues are now being addressed by the proponents of Resource Description and Access, or RDA, an initiative undertaken by national libraries and standards agencies that develop recommendations for library cataloging (JSC-RDA 2006). The goal of RDA is to replace the semantic standard most prevalent in the library community, the Anglo-American Cataloging Rules, with a more modern standard that recommends "a consistent data encoding for elements that matter the most" (Delsey 2009). Two changes proposed by RDA supporters are especially relevant to the ONIX-MARC relationship and the problems I have described in this article.

First, some descriptions are simply revamped. For example, a draft version of an RDA vocabulary for physical descriptions suitable for both ONIX and MARC has already been made public (Dunsire 2007, JSC 2006b). If this effort continues, only the structural manipulation required to create the appropriate output syntax would be required because the semantics would be exactly the same. But this a long-term vision because much difficult work remains to be done.

The second long-term vision is to treat many of the elements in a bibliographic description as linked data (Heath 2009 and Hillmann, et al. 2010), a change that would also greatly simplify the relationship between ONIX and MARC because data would no longer be copied when records are mapped. Instead, records conforming to the two standards would simply point to the same authoritative resources for contributor names, titles of works, standard identifiers, and subjects, thus eliminating all potential for typographical errors and slight shifts in meaning or granularity of description I have noted in this article.

What remains in the program of semantic harmonization is some cleanup that is not directly addressed by the goals of RDA. For example, constant vigilance is required to ensure that ONIX and MARC codes used to identify the types of supplementary material, the nature of a contributor's work, or the characteristics of physical media and carriers are, if not identical, at least equivalent in scope and descriptive power. Nor does RDA address the problem of how to encode the sometimes lengthy text of supplementary materials. But since a table of contents or a review has the same meaning in the two standards, stakeholders simply need to design an unambiguous markup scheme that satisfies the requirements of both communities.

There is an ongoing need to manage pragmatic differences between the two standards.

Semantic harmonization of ONIX and MARC will be a major step forward, but the work won't be finished because there will always be differences in how libraries and publishers use bibliographic data. Moreover, pragmatic need is the primary driver for change. Going forward, this change will have to be managed with a heightened awareness of the impact it will have on all affected communities of practice.

References

All URLs were accessed on March 18, 2010.

BISG (Book Industry Study Group). 2005. Product metadata best practices for data senders. http://www.bisg.org/docs/Best_Practices_Document.pdf.

———. 2009. BISAC subject headings list, major subjects – 2009 edition. http://www.bisg.org/what-we-do-0-136-bisac-subject-headings-list-major-subjects---2009-edition.php.

Chan, Lois Mai. 2007. *Cataloging and classification: An introduction*. Lanham, Md: Scarecrow Press.

Delsey, Tom. 2009. Look before you leap: RDA compared to AACR2. Presentation given at the American Library Association Midwinter meeting, January 10, 2009, Denver, Colorado, USA. http://presentations.ala.org/images/1/10/LLL-Delsey-ALA2009.ppt.

Dunsire, Gordon. 2007. Distinguishing content from carrier. *DLIB Magazine*, 13:12. http://www.dlib.org/dlib/january07/dunsire/01dunsire.html.

EDItEUR. 2009a. ONIX releases prior to 2.1. http://www.editeur.org/95/Releases_prior_to_2.1_rev03#Release%202.1%20rev%2002.

———. 2009b. ONIX: About Release 3.0. http://www.editeur.org/12/About-Release-3.0/.

Godby, Carol Jean, Devon Smith and Eric Childress. 2008a. Toward element-level interoperability in bibliographic metadata. *Code4Lib Journal*, Issue 2. http://journal.code4lib.org/articles/54.

———. 2008b. Encoding application profiles in a computational model of the crosswalk. *International Conference on Dublin Core and Metadata Applications, DC-2008*. http://dcpapers.dublincore.org/ojs/pubs/article/viewArticle/914.

Heath, Tom. 2009. Linked data: Connect distributed data across the web. http://linkeddata.org/.

Hillmann, Diane, Karen Coyle, Jon Phipps, and Gordon Dunsire. 2010. RDA vocabularies: Process, outcome, use. *DLib Magazine*, 16:1/2. http://www.dlib.org/dlib/january10/hillmann/01hillmann.html.

IFLA. 2007. International standard bibliographic description. http://www.ifap.ru/library/book264.pdf.

———. 2010. Functional requirements for bibliographic records. http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records.

JSC-AACR2 (Joint Steering Committee for Revision of AACR). 1988. *Anglo-American cataloging rules, second edition. Revisions 1985.* Chicago: American Library Association.

JSC-RDA (Joint Steering Committee for the Development of RDA). 2006a. RDA: Resource description and access. http://www.rda-jsc.org/rda.html.

———. 2006b. RDA/ONIX framework for resource categorization. http://www.loc.gov/marc/marbi/2007/5chair10.pdf.

Library of Congress. 2000. MARC 21 record builder. *Network Development MARC Standards Office*. http://www.loc.gov/marc/marc2onix.html.

———. 1999. MARC 21 format for bibliographic data. *Network Development MARC Standards Office*. http://www.loc.gov/marc/bibliographic/.

———. 2009. What is a MARC record, and why is it important? http://www.loc.gov/marc/umb/um01to06.html.

OCLC. 2009. OCLC metadata services for publishers. http://publishers.oclc.org/en/metadata/default.htm.

———. 2010a. ONIX-MARC mapping. Excel spreadsheet. http://www.oclc.org/research/publications/library/2010/2010-14a.xls.

———. 2010b. WorldCat. http://www.worldcat.org/.

RDA (Resource Description and Access). 2010. Constituency review. http://www.ifap.ru/library/book264.pdf.