# From Records to Streams: Merging Library and Publisher Metadata

Carol Jean Godby
OCLC, USA
godby@oclc.org

## Abstract

This article announces the availability of a crosswalk between ONIX 2.1 and MARC 21 developed by OCLC and illustrates how it is used in the *OCLC Metadata for Services Publishers* project. To accomplish the goal of merging library and publisher metadata and anticipating the need to mine MARC records for other purposes, the design of the crosswalk, the corresponding software, and the application take records apart and process the fields individually, creating data streams that match the intended use of the ONIX standard. Though this design works well enough to support commercial-grade processes, problems arise with mappings between physical descriptions in the two standards, which need to be more rigorously modeled or closely aligned. Nevertheless, the RDA/ONIX Framework, which is reviewed here, promises to reduce this obstacle.

**Keywords:** crosswalks; interoperability; library metadata; MARC; ONIX; publisher supply chain; RDA; Resource Description and Acces.

## 1. A crosswalk from ONIX to MARC

ONIX, or ONline Information eXchange, is a set of international standards introduced by EDItEUR[1] in 1997 for the conduct of e-commerce in the publishing community. ONIX for Books is the most widely used. With an XML syntax, extensive use of coded data, and a record structure that supports many applications, an ONIX record tracks a book through the supply chain as it is developed, published, registered by national libraries, shipped to distribution centers, marketed, sold, translated, and re-introduced in alternative formats. Along the way, the ONIX record acquires a more and more detailed bibliographic description and is constantly updated with status information that lets interested parties know where the item can be procured, how much it costs, and whether it has access restrictions. Supporting organizations such the New York-based Book Industry Study Group, or BISG, and the UK-based Book Industry Communication, or BIC, recommend best practices for the creation of ONIX records, conduct seminars, develop subject heading schemes, and coordinate interaction with related communities.

Recognizing the need for closer relationships with the publishing community, research and development staff at OCLC have spent the past two and a half years studying the landscape of library and publisher metadata. One outcome is a recently published a crosswalk, or set of semantic mappings, for Version 2.1 of ONIX for Books and MARC 21 (OCLC, 2010a), which represents a major upgrade to the publicly available statements about the relationship between these two standards. Work is underway to create a crosswalk to MARC for ONIX for Books 3.0, which will be available in the final quarter of 2010.

The article 'Mapping ONIX to MARC' (Godby, 2010) highlights some of the key differences between the two standards and acts as an introduction to the crosswalk. At the highest level, the crosswalk is represented as a human-readable Excel spreadsheet containing thirteen worksheets, or tables. The most important is the table labeled 'ONIX,' which captures the high-level relationships between the ONIX elements required for a bibliographic description and their closest MARC counterparts. The rest of the tables are subordinate to this goal and specify details

[1] EDItEUR. (2010). http://www.editeur.org/.

such as maps between controlled codes or complex conditional relationships involving these values.

In OCLC's implementation, the published crosswalk is refactored slightly so it can be processed algorithmically to build one fully populated MARC field at a time from the ONIX source.[2] The translated fields are then packaged into one of several output syntaxes, such as MARC 2709, MARC-XML, or a locally designed structure. As described in more technical detail elsewhere (Godby, et al. 2008a and 2008b), a spreadsheet with separate columns that label the source, target, translation conditions, and operations on data values is submitted to an automated process that generates executable code in a custom-designed scripting language that closely models the concepts in a crosswalk. For example, a metadata standards expert who maps ONIX <TitleText> to MARC 245 $a asserts that both statements are essentially equivalent ways of expressing the 'title' concept. As specified in Rows 42-48 in the worksheet labeled 'ONIX,' primary titles are mapped to 245 $a, while translations are mapped to 247 $a and alternative titles are mapped to 246 $a. In addition, the $h subfield must be populated with values triggered by codes in the ONIX <ProductForm> element, which are described in Table 2 of the crosswalk.

The map between the ONIX and MARC Title elements illustrates some of the complexity captured in the crosswalk. The map between title elements succeeds because the concepts are semantically similar, though not identical because the MARC 245 field is supplemented with information derived from fields outside the ONIX <Title> composite to ensure that it can serve as an access point to facilitate retrieval from a library catalog, a concept that is alien to the ONIX standard. But even with this added layer of meaning, the ONIX and MARC Title elements are reasonably transparent and independent of the other elements in the record. As a result, the title data can be extracted from its source and applied to a new data stream, complementing the maps for contributors, subjects, publishers, identifiers, and all other elements that make up a bibliographic description. According to the framework described by Marcia Zeng and Lois Mai Chan (Chan and Zeng 2006; Zeng and Chan 2006), this one-by-one mapping of elements from different metadata schemas accomplishes interoperability at the element level, which must occur if the higher-order interoperability of schemas, records, and repositories is to be achieved.

Interoperable records produced through metadata crosswalks support multiple uses across different communities of practice by making it possible to convert or access records in different standards and track the supply and demand of corresponding physical materials. Thus the ONIX to MARC crosswalk presents a real-world opportunity to test design ideas about how to define relationships and manage change, and in so doing, leverage some of the library community's investment in metadata.

## 2. Streams of library and publisher metadata

As described in Library of Congress documentation (LC, 2006), a well-formed MARC record has three dimensions. First, it has *record structure*, or a syntax derived from international standards such as *Format for Information Exchange* (ISO 2709) or *Bibliographic Information Interchange* (ANSI/NISO Z39.2), the American counterpart. Second the MARC record has a *content designation*, or a set of tags, subfields, indicators, and coded values required for a bibliographic description. Finally, the MARC record has *content* that can be formally validated by making reference to cataloging rules such as AACR2 and formatting standards such as the International Standard Bibliographic Description (ISBD). A crosswalk is defined for the content designation, which has easily identifiable counterparts in other standards such as ONIX. The record structure is also relevant because translation software that invokes a crosswalk must read and produce the correct syntax for the source and target standards. But the MARC record content dimension is problematic and perhaps out of scope of the metadata crosswalking enterprise because it is tightly

[2] Table 3, discussed later in this article in another context, shows a small example of the design that was adopted.

bound to the single use case of supporting discovery and retrieval from library catalogs, which exhibits considerable local and regional variation. By contrast, the ONIX standard makes no assumptions about how the data will be used, or even whether the record is a primary concept. Though an ONIX record can be constructed for a temporary expediency, requirements in the publisher supply chain place greater value on the ability to add, subtract, modify, or rearrange the individual data elements that comprise the record. How can successfully can the information locked inside MARC records be extracted and merged with streams of ONIX data? And what use cases can be supported?

### 2.1. Metadata for libraries and publishers

These questions are being explored in the *OCLC Metadata Services for Publishers* project (OCLC, 2009), whose goal is to streamline the relationship between library and publisher metadata. Libraries, retailers, wholesalers and aggregators are consumers of publisher direct and publisher supply chain metadata, many of which conform to one of the ONIX standards. Conversely, some parts of the publisher supply chain use and create library metadata, which are typically MARC records. Since libraries buy most of their materials from publishers or associated vendors, they consult ONIX data for discovery and procurement, triggering a cascade of supporting tasks. For example, a cataloging-in-publication MARC record is produced at the Library of Congress from drafts of publisher-supplied descriptions. At the library, buying decisions might require a database search for the same or similar items, which could involve a software process that performs a match between ISBNs or other identifiers found in ONIX and MARC records. Once the item is obtained, it must be represented by a MARC record in a local database, which requires a translation of an ONIX record. Though publishers and wholesalers perform some of these tasks, they benefit from the investment of effort because the extra attention given to names and subject headings to accommodate the needs of libraries also makes published products easier to discover by non-library customers.

It is tempting to point out that libraries and publishers represent two communities of practice, which have developed two metadata standards that represent essentially the same information. But the overlap between the standards also masks a productive division of labor. Publishers are focused on what is moving through the marketplace right now: who is responsible for making it available; how it is priced; where it can be obtained; whether its access is restricted by intellectual property rights; and which versions, physical formats, and editions are available. Of course, libraries also need to be concerned about these issues because they are among the publishers' customers, but they are engaged in the fundamentally different mission of preserving materials for access and use by patrons in perpetuity. Performing this mission requires the creation of archival-quality descriptions with authority-controlled forms of names and subject headings. This long-range view also benefits from the development of innovative models of bibliographic description such as Functional Requirements for Bibliographic Records, or FRBR (IFLA, 2010), which establish relationships among related versions of works. Since both communities continue to innovate—as librarians merge authority files across international boundaries in projects such as the Virtual International Authority File (VIAF, 2010), or as publishers grapple with the anticipated widespread adoption of e-books—it will become even more important to merge the two streams of metadata at the earliest point and continuously update them. And the crosswalk is the most important piece of infrastructure for making this transform happen.

### 2.2. A process flow.

Figure 1 depicts the process flow developed for the initial release of the *OCLC Metadata Services for Publishers* project. ONIX records obtained from publishers (1) are translated to MARC (2) using the crosswalk described in this article. The translated MARC record is matched with a set of records representing the same intellectual work, using fuzzy matching algorithms against a version of OCLC's WorldCat database to which a FRBR-clustering algorithm has been applied. In the enrichment step (3), fields from the FRBR cluster are applied to the record, which is

translated (4) to MARC (5) and to ONIX (6), using logic that is similar to the ONIX-to-MARC crosswalk.

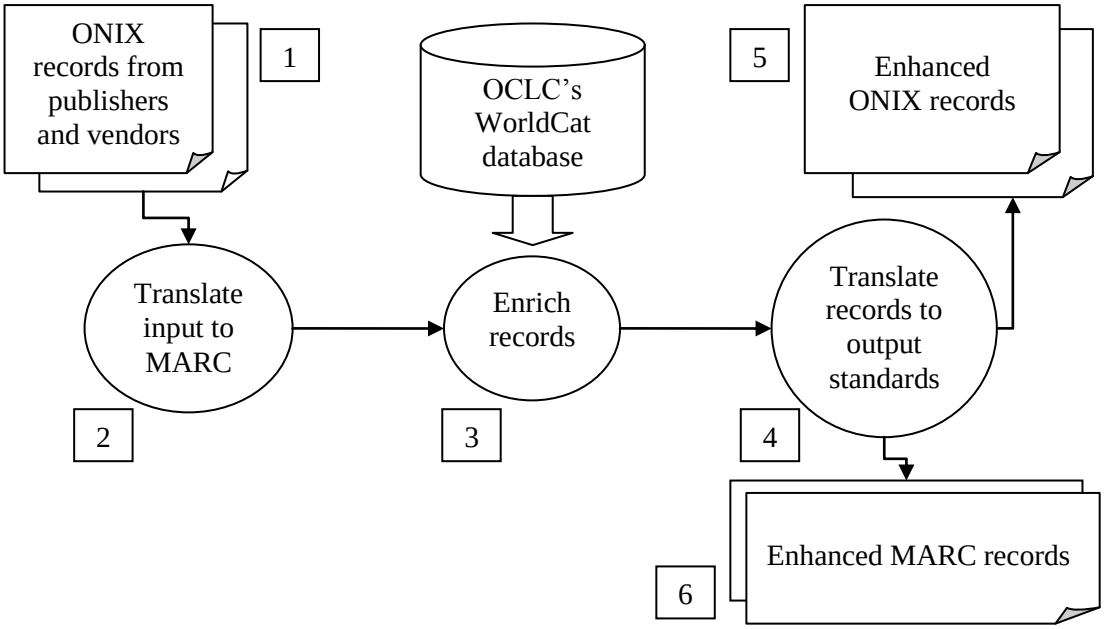


FIG. 1. Data flow through the *OCLC Metadata Services for Publishers* system.

Below, Table 1 shows a result for a paperback book on primary education. Here, an exact match has been found between the corresponding ONIX and MARC records, using the ISBN number as a key. Information from the ONIX record that maps to MARC is shown in the white cells in the table. Note that the ONIX data matches the name and spelling of the publisher and author as well as the prices in two currencies, all of which constitute important verification of the library data because publishers are the ultimate authority on their own names and the transaction data for the products they manufacture. Information obtained from the MARC record is shown in the shaded boxes. Added fields include a Library of Congress Control Number, Library of Congress and Dewey call numbers, subject headings, and notes. These results reflect the complementary strengths of the two communities that supplied the metadata. In general, libraries contribute subject headings and authority control on titles, subjects, contributors, while publishers contribute authoritative data pertaining to the physical aspects of their products, transaction and status information, and copyright status.

TABLE 1. An ONIX record enhanced with MARC fields.

| **Data element** | **Original ONIX data** | **Enhanced ONIX data** |
|---|---|---|
| LCCN | | 2005057403 |
| ISBN 10 | 0414394279 | 0414394279 |
| LC call number | | LB1031 |
| LC call number | | .D35 2006 |
| BISAC subject code | EDU | EDU |
| BISAC subject code | 0000 | 0000 |
| DDC edition | | 22 |
| DDC number | | 372.13/94 |
| Author | Dean, Joan | Dean, Joan |

| Title | Meeting the Learning Needs of All Children: Personalised Learning in the Primary School | Meeting the Learning Needs of All Children |
|---|---|---|
| Subtitle | | Personalised Learning in the Primary School |
| Place of publication | | London; New York |
| Publisher/Imprint | Routledge | Routledge |
| Date of publication | 2006 | 2006 |
| Note | | Includes bibliographical references (p. [87]) and index. |
| LCSH | | Individualized instruction; Education, Elementary; Multicultural education; Ensignement individualisé; Ensignement primare; Éducation interculturelle |
| LCSH | | Great Britain; Grande Bretagne |
| Price/Currency | 31.95/USD | 31.95/USD |

This record illustrates a proof-of-concept for the proposal that metadata can be improved for both libraries and publishers by a software process featuring crosswalks between ONIX and MARC that provide access to a sophisticated enrichment module. In day-to-day operations, such a process might help publishers cost-effectively create rich records for backlist or out-of-print materials that have only minimal descriptions containing little more than titles, authors, and ISBNs. Enhanced records can also be created more indirectly. If the description represents a new edition or a physical format with an ISBN that does not return an exact match in WorldCat, the record can still be populated by fields applied from other records in the same FRBR cluster. Given that the mapping and enrichment processes are fully automated, they can be applied again and again as more evaluative content and ancillary links are made available from publishers, and as innovations in bibliographic description are implemented by the library community.

For the most part, the software modules implemented in the OCLC Metadata Services for Publishers project operate on individual metadata elements in the two data streams. For example, equivalences between the ONIX and MARC elements shown in Table 1 are mapped, preparing the incoming ONIX record for enhancement; corresponding operations are performed on the output. Fields from the matching database records are applied. If there is a discrepancy between two fields with the same information, an automated process determines which field is more reliable or useful, such as the MARC field shown in Table 1 that has more detailed coding for the title because it labels the subtitle separately. Misspelled data in subject or contributor fields can also be detected and replaced with authority-controlled versions. Once all relevant fields have been added to the input record, other software processes detect and remove redundant information. Only at this point is the output submitted to record-level analyses. After the record is mapped back to ONIX for delivery to publishers, it is validated against EDItEUR's published XML schema and checked for compliance with best practices coding standards defined by the Book Industry Study Group. The new MARC record is validated for conformance to AACR2 rules before it is added to OCLC's WorldCat database.

## 3. Mapping physical descriptions

Execution of the tasks in the *OCLC Metadata Services for Publishers* project requires that old and new paradigms for metadata representation be coordinated. The thirteen-year-old ONIX standard is element-oriented, designed for transmission, and agnostic about how the data will be used. But the forty-one-year-old MARC standard is record-oriented, designed for storage in

databases, and tailored to the needs of particular applications in the library community. To the extent that projects like the one I have described are successful, these differences are immaterial. But problems are sometimes introduced by AACR2 semantics and ISBD punctuation, which are required for semantically correct bibliographic records and hint at the restricted context for the expected use of the MARC standard. These levels of encoding go beyond what is required to transmit isolated elements such as identifiers, titles, and subjects make the maps between ONIX and MARC difficult to maintain. If left unresolved, these problems may result in loss of information. The reader is referred to Godby (2010) for the big picture, but in the rest of this article I want to concentrate on the worst problem I identified in that study of mappings from ONIX to MARC and describe the issues in more depth than I could there.

### 3.1. What is a physical description?

In the publisher as well as the library supply chain, it is fundamentally important to describe the physical characteristics of an item. Now that so much information is consumed through electronic media, the patron may even not be able to access the desired content if a Blu-Ray disc is ordered but a high-density DVD is delivered, or if the Amazon Kindle e-book format is delivered instead of the Sony or iPad format. To ensure that the request is correctly fulfilled, the essential characteristics of the physical item—and, where appropriate, the electronic mediation device—must be described. Yet this facet of the bibliographic description is complex, opaque, and brittle in the current ONIX-to-MARC crosswalk.

Figure 2 shows fragments of matching ONIX and MARC records generated from the OCLC crosswalk that describe a compact disc containing selected recordings of arias from Puccini's operas by various artists. The most important element in the ONIX record is 'AC,' the value of <ProductForm> indicating that the record describes an audio compact disc. The values that are mapped from ONIX to MARC are shown in bold and the values that are copied are shown in italics. As the *ProductForm* worksheet in the crosswalk shows, the ONIX <ProductForm> value of 'AC' triggers a one-to-many map to six MARC fields—one of which, the MARC 007 field, required by AACR2 cataloging rules for non-book media, 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. In addition, the Leader/01 value of 'd' marks this as a sound recording, a designation that is repeated in the 245 $h subfield, which contains the keyword 'sound recording.'

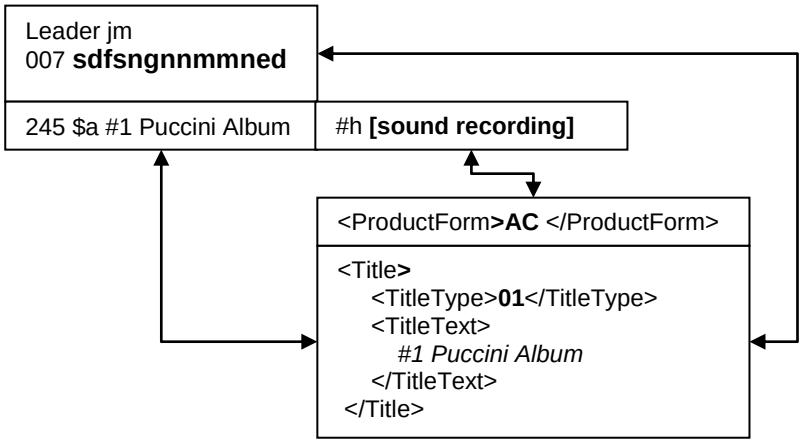


FIG. 2. ONIX <ProductForm> 'AC' and a corresponding AACR2-encoded MARC description.

Note that the end-user term 'CD' appears in neither the MARC nor the ONIX description. 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 *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 merely 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 specification will surely change over time and they could even now be subject to regional differences. If so, this map would have to be replicated many times, with different values for some of the 007 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.

### 3.2. Print or digital?

Unfortunately, the map between the MARC and ONIX physical description elements given in the table labeled *ProductForm* in the crosswalk understates the true complexity of the relationship. The additional descriptive burden can be traced to the fact that ONIX makes a fundamental distinction between printed books and materials such as CDs and DVDs, whose content must be experienced through an electronic device. If the item is a book, the bibliographic description contains a page count; if the item is a storage device for an electronic medium, the description should have an <Extent> element, which specifies a running time or file size. And both classes of materials can have item counts and physical dimensions. Since MARC does not make the same distinctions as ONIX, all of this information must be represented in the overloaded 300 field. As defined in the MARC standard (LC, 1999), *extent* is the "number of physical pages, volumes, cassettes, total playing time, etc." Thus the ONIX data values in <NumberOfPages>, <NumberOfPieces>, and <Extent> map to MARC 300 $a, while physical dimensions are mapped to $c. The distinctions are easily lost because explicitly labeled data in ONIX is converted to text literals in MARC, all of which are difficult to process algorithmically. Figure 3 shows another fragment of the record for the audio CD described above with the results of these maps.

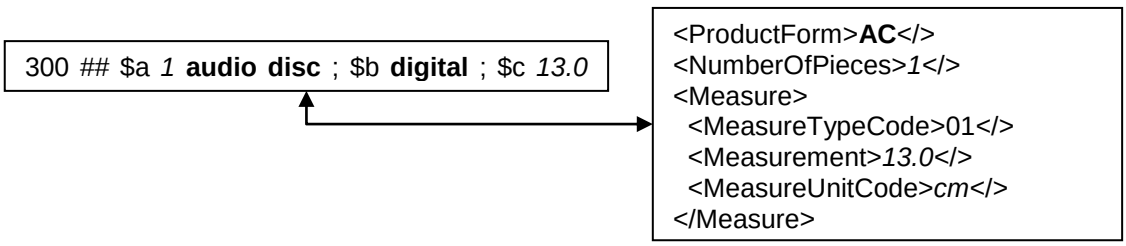


FIG. 3. ONIX sources for the MARC 300 field.

In short, the relationship between the ONIX and MARC physical description elements is many-to-many and difficult to comprehend, a symptom of an underlying conceptual problem.

## 4. Does RDA solve the problem?

Fortunately, the problems with physical descriptions in ONIX and MARC are being addressed by the proponents of Resource Description and Access, or RDA, a new cataloging standard sponsored by national libraries and professional organizations in the English-speaking world (RDA, 2010). In particular, a task force reporting to the RDA Joint Steering Committee is developing a common vocabulary for physical description elements that aligns the concepts in ONIX and MARC more closely (Dunsire, 2007). One outcome is the first draft of the *RDA/ONIX*

*Framework for Resource Categorization* (Kiorgaard, 2006). The vocabulary defined in the appendices of this report has been registered in the RDA repository maintained by Hillmann, et al. (2010). When RDA replaces AACR2 as the encoding scheme of choice for MARC records in Anglophone communities, some of the problems I have described will be mitigated, but it is instructive to work out in detail just how the mappings from ONIX to MARC will be affected.

### 4.1. Content, Media, and Carrier

One essential relationship is defined in Appendix B of the *RDA/ONIX Framework* for the concepts *StorageMediumFormat*, *HousingFormat*, and *IntermediationTool*. All of these are attributes of *Resource Carriers*, one of three newly defined RDA categories that can be used instead of the combination of Leader, 007, and 008 values now used to describe the physical object in an AACR2-encoded MARC record. The two companion concepts are *Media Type* and *Content Type*. As spelled out in the table in Appendix D of the *Framework*, an audio disc is a resource carrier that has a StorageMediumFormat value of 'disc,' a HousingFormat value of 'not applicable,' and an IntermediationTool value of 'audio player.' Values for other types of carriers are shown in Table 2, which is an excerpt from the Appendix D table.

TABLE 2. Carrier types in the RDA/ONIX framework.

| BaseCarrierCategory | StorageMediumFormat | | | | | | | | HousingFormat | | | | | | IntermediationTool | | | | | | | | Sample Category Label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | sheet | strip | roll | disc | sphere | cylinder | chip | file server | binding | flipchart | reel | cartridge | cassette | not applicable | microform reader | microscope | projector | stereoscope | audio player | audiovisual player | computer | not required | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | |
| BaseCarrierCategory 3:5:1 | | | ■ | | | | | | | | | | ■ | | ■ | | | | | | | | microform cassette |
| BaseCarrierCategory 3:5:3 | | | ■ | | | | | | | | | | ■ | | | | ■ | | | | | | film cassette |
| BaseCarrierCategory 3:5:5 | | | ■ | | | | | | | | | | ■ | | | | | | ■ | | | | audiocassette |
| BaseCarrierCategory 3:5:6 | | | ■ | | | | | | | | | | ■ | | | | | | | ■ | | | videocassette |
| BaseCarrierCategory 3:5:7 | | | ■ | | | | | | | | | | ■ | | | | | | | | ■ | | computer cassette |
| BaseCarrierCategory 3:6:3 | | | ■ | | | | | | | | | | | ■ | | | ■ | | | | | | filmstrip |
| BaseCarrierCategory 3:6:5 | | | ■ | | | | | | | | | | | ■ | | | | | ■ | | | | audio roll |
| BaseCarrierCategory 4:4:7 | | | | ■ | | | | | | | | ■ | | | | | | | | | ■ | | computer disc cartridge |
| BaseCarrierCategory 4:6:4 | | | | ■ | | | | | | | | | | ■ | | | | ■ | | | | | stereograph reel |
| BaseCarrierCategory 4:6:5 | | | | ■ | | | | | | | | | | ■ | | | | | ■ | | | | audiodisc |

The terms listed in the final column of Table 2, *Sample Category Label*, provide the link to third-party metadata schemas such as ONIX—in particular, Codelist 7, whose values populate the <ProductForm> element. But the mapping is unreliable because neither the glosses in Codelist 7 nor the labels in the RDA table are controlled terms. Nevertheless, as Dunsire (2007) suggests, it is possible to make some informal associations using this information. For example, the ONIX code DH from Codelist 7, which is glossed as 'online resource,' is defined by a StorageMediumFormat value of 'file server,' a HousingFormat value of 'not applicable,' and an IntermediationTool value of 'computer.' But Dunsire emphasizes that this is only an informal example. Much more input is required from the RDA and ONIX communities to establish more definitive relationships. One of the most important issues yet to be addressed is the difference in granularity between Codelist 7, which has upwards of 100 entries (and continues to grow), and the Appendix D table, which has only fourteen entries. Significantly, Appendix D is not granular enough to resolve the mapping between ONIX and MARC shown in Figures 2 and 3 because it cannot distinguish between CDs, which have a <ProductForm> value of AC, and non-CD audio discs, which have a <ProductForm> value of AE. Only the basic categories are represented in this table, though the vocabulary scheme is designed to be extensible through the definition of qualified categories, which are defined by stakeholder communities and are not part of the framework.

### 4.2. Mapping RDA from ONIX.

Now consider the relationships shown in Figure 4, which shows the effects of mapping from ONIX to an RDA-encoded MARC record. On the left is a subset of the critical fields of the same ONIX record shown in Figures 2 and 3; on the right is the reformulated MARC record. This example is based on a sample record provided by Tillett (2009). Note the three new fields. The 336 field is the *RDA Content* type, which is broadly equivalent to the AACR2 meaning of Leader/06, or 'Type of record', and is defined as the "fundamental form of communication in which the content is expressed" (RDA, 2010, ch. 6.10). The 337 field designates the *RDA Media* type, which corresponds to 007/00. And the 338 field is the *RDA Carrier* type, which is equivalent to Leader/01 (Specific material designation), or the type of storage medium required to experience the content. The 300 field has essentially the same information as the original AACR2 record.

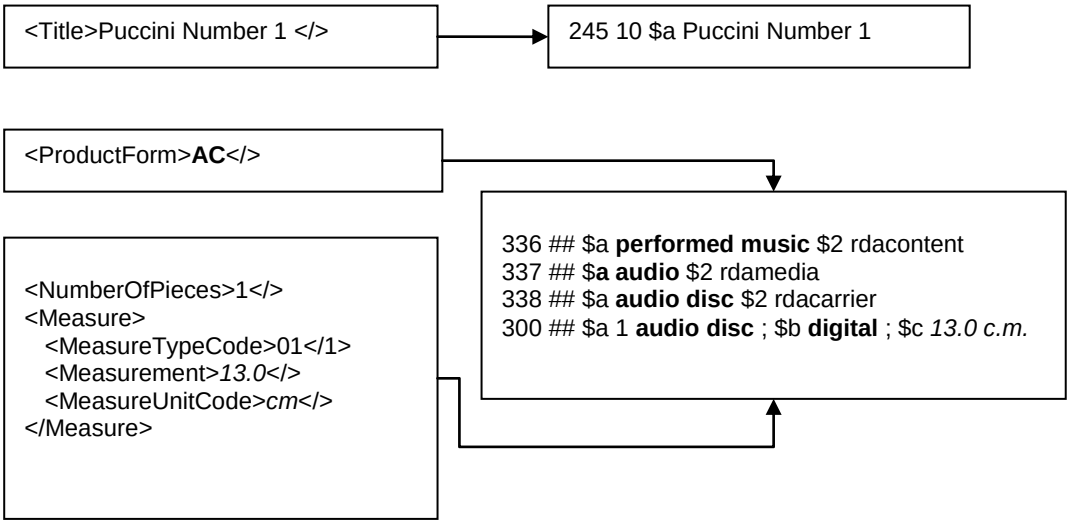


FIG. 4. ONIX physical descriptors and corresponding RDA-encoded MARC.

What is the effect of the RDA encoding on the relationship between ONIX and MARC? At first glance, it appears that RDA produces a slightly simpler and arguably more transparent relationship because the shared vocabulary eliminates the need to map to most of the 007 subfields. This has the beneficial side effect of reducing the problem of overspecification discussed above. But it is still a complex and brittle many-many mapping. In Figure 4, the RDA equivalents of ONIX <ProductForm>AC</>, shown in bold fonts in the MARC record, populate the 336-338 fields and $a and $b of the 300 field, while values in the <Measure> composite, shown in italic fonts, are mapped to 300 $c. But one of the major problems identified in the AACR2 encoding remains. Since the text strings in the 300 field are built up from uncontrolled vocabulary and can assume multiple forms, the map to ONIX must account for this variability. Mapping failures will occur if key values in the input text have unexpected spelling, formatting, or typographical errors.

The same point can be made about the maps for the newly defined elements. In the AACR2 encoding, the equivalent information in the 337 and 338 fields is mapped to controlled codes in the 007 field; in the proposed revision, it is mapped to text fields for which RDA proponents make recommendations but also permit some variation. Moreover, the closest semantic match to the ONIX <ProductForm> element is 338, which contains the rdacarrier value. But in the currently available draft of the registered vocabulary, the permissible values for $a are not precise enough to make critical distinctions between an audio CDs and a non-CD audio disc, which is critical for the record excerpted in Figures 2 and 3. Likewise, the vocabulary cannot distinguish among various forms of videodiscs or most other carriers. To ensure a successful mapping, the

distinguishing characteristics of the formats in question must be recorded in the 300 $a and $b text fields (and consulted when the MARC record is mapped to ONIX). The result is a one-to-many mapping from the ONIX <ProductForm> code to a MARC target that contains a mixture of coded and textual values. To make the mapping more robust, metadata experts in the library and publishing communities need to expand the registered carrier vocabulary. But when it becomes available, the sponsors of the RDA best practices guidelines need to determine whether the descriptors in the 300 field are superfluous, and if so, adjust their recommendations for how to use them.

The implied translation from an ONIX source to an RDA-encoded MARC record shown in Figure 4 is shown explictly in the crosswalk fragment in Table 3 below. Though the information is represented redundantly, this table shows the relevant relationships in a form that could be converted automatically to executable code using software such as OCLC's Crosswalk Web Service (Godby, et al., 2008). Each map generates a complete MARC field. The first three maps generate the newly defined 33x fields and Map 4 generates the 300 field. But Map 5 is also required because backward compatibility with AACR2 permits the RDA media, content, and carrier descriptors to appear in the Leader and 007 fields (RDA, 2010, ch. 6).

TABLE 3. Some maps from ONIX to RDA-encoded MARC.

| **SOURCE** | | | | **TARGET** | |
|---|---|---|---|---|---|
| *Map* | *ONIX composite* | *ONIX element* | *Value* | *MARC field* | *MARC subfield* |
| 1 | | ProductForm | AC | 336 | $a= performed music |
| | | | | | $2=racontent |
| 2 | | ProductForm | AC | 337 | $a=audio |
| | | | | | $2=rdamedia |
| 3 | | ProductForm | AC | 338 | $a=audio disc |
| | | | | | $b=digital |
| | | | | | $2=rdacarrier |
| 4 | | ProductForm + NumberOfPieces | AC | 300 | $a= NumberOfPieces + audio disc |
| | | | | | $b=digital |
| | Measure | Measurement + MeasureUnitCode | | 300 | $c |
| 5 | | ProductForm | AC | Leader | 01=j |
| | | | | 007 | 00=s |
| | | | | 007 | 01=d |

In sum, the RDA/ONIX framework for resource categorization makes a first step toward solving the problems that arise when physical description elements are mapped from ONIX to MARC. But the resulting crosswalk is, at best, only somewhat less problematic than the version that assumes an AACR2 encoding because the common vocabulary does not produce transparent one-to-one mappings, textual values are substituted for more strictly controlled coded values, and the complexity of the relationships is only slightly reduced. Much work remains to be done by relevant stakeholders to solidify the strategically important relationship between these two standards. First, a concerted effort must be made to model difficult domains such as physical descriptions and define common vocabulary because transparent mappings will not be possible otherwise. Indeed, a detailed look at this problematic facet of the bibliographic description sheds

light on why AACR2 and, perhaps also RDA, will continue to require record-level validation and other macro-operations that will pose problems for any effort to access, extract, map, and update individual MARC elements: information about some key concepts is inexplicably scattered throughout the record. Second, proponents need to expedite the implementation one of the long-term promises of RDA, which is to replace error-prone textual data values with linked data (Hillmann, et al., 2010), which will make authoritative sources of descriptive vocabulary, names, and subject headings available to all participating standards. These outcomes will eliminate most of the problems with physical descriptions I have identified in the previous discussion. They will also improve many of the other mappings from ONIX to MARC and make it easier to establish reliable connections other metadata standards, such as Dublin Core Terms or MODS.

## 5. Future directions

The failure to map physical descriptions from ONIX to MARC is potentially catastrophic in supply-chain transactions because a mismatch between the physical format of the item and the available electronic mediation device could make the content inaccessible even if the customer or library patron successfully obtains a requested item. But the issues identified here also appear in most of the other maps of elements in a bibliographic description in ONIX and MARC. First, most of the maps involve text and must account for differences in punctuation and formatting not present in coded data, whose values are more strictly controlled and verifiable by semantic validity checks. Second, the two standards evolve at different rates, often requiring burdensome changes to already complex and opaque relationships. Finally, the two standards have different philosophies about backward compatibility. Work is underway to develop a fundamentally new crosswalk to ONIX 3.0 because this version is not backwardly compatible with ONIX 2.1, which is represented in the publicly accessible crosswalk developed at OCLC. But, as we have seen, the proponents of RDA have made a different decision because some of the changes to MARC required to accommodate RDA must be compatible with previous versions, which would add complexity to crosswalks involving the two versions of ONIX.

These observations are consistent with those made in a recent anthology of studies of MARC tag usage in OCLC's WorldCat database (Smith-Yoshimura, et al. 2010). Examining the discovery of bibliographic descriptions in library catalogs, the typical application for MARC records, as well as matching, linking, collection analysis, and record conversion, the authors noted that only a small subset of the fields defined in the standard are involved in machine processing, while many others are used inconsistently from one installation to the next or contain textual values that cannot be easily manipulated or interpreted. These automated processes are also hampered by redundant information and content that is split across multiple fields. Such problems imply, according to the authors, that MARC is "a niche communication format approaching the end of its life cycle" (Smith-Yoshimura, et al. 2010, p. 14). But even if the library community moves toward a more modern standard, there will be a need for robust crosswalks to ingest the hundreds of millions of legacy records created in the library community and mine the knowledge contained in them. This process will require a metadata model that extracts elements one at a time and recombines them, putting them to previously unanticipated uses, much as ONIX elements are deployed now to process transactions in the publisher supply chain.

## Acknowledgements

I would like to acknowledge the contributions of my colleagues Bob Pearson, who developed the crosswalk, and Renee Register, who manages *OCLC Metadata Services for Publishers*. They are full collaborators in the work described here, but they are not responsible for any errors in my writeup.

## References

Chan, L M. & Lei Zeng, M.L. (2006). Metadata interoperability and standardization—A Study of methodology part I: Achieving interoperability at the schema level. *DLIB Magazine*, Volume 12, Number 6. http://www.dlib.org/dlib/june06/chan/06chan.html.

Dunsire, G. (2007). Distinguishing content from carrier: The RDA/ONIX framework for resource categorization. *DLIB Magazine*, Volume 13, Number 12. http://www.dlib.org/dlib/january07/dunsire/01dunsire.html.

EDItEUR. (2010). http://www.editeur.org/.

Godby, C. J. (2010). Mapping ONIX to MARC. Report produced by OCLC Research. Published online at: http://www.oclc.org/research/publications/library/2010/2010-14.pdf.

Godby, C. J., Smith D. & Childress, E.. (2008a). Toward element-level interoperability in bibliographic metadata. *Code4Lib Journal*, Issue 2. http://journal.code4lib.org/articles/54.

Godby, C. J., Smith, D. & Childress. E. (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.

Hillmann, D., Coyle, K., Phipps, J., and Dunsire, G. 2010. RDA vocabularies: Process, outcome, use. *DLib Magazine*, 16:1/2. http://www.dlib.org/dlib/january10/hillmann/01hillmann.html.

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

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

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

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

Kiorgaard, D. (2006). RDA/ONIX framework for resource categorization. http://www.loc.gov/marc/marbi/2007/5chair10.pdf.

LC (Library of Congress). (2006). MARC 21 format for bibliographic data: Introduction. http://www.loc.gov/marc/bibliographic/bdintro.html.

RDA (Resource Description and Access). (2010). Constituency review. http://www.rdatoolkit.org/constituencyreview.

Smith-Yoshimura, K., Argus, C., Dickey, T. J., Naun, C. C., Rowlinson de Ortiz, L., & Taylor, H.. (2010). Implications of MARC tag usage on library metadata practices. Report produced by OCLC Research in support of the RLG Partnership. Published online at: http://www.oclc.org/research/publications/library/2010/2010-06.pdf.

Tillett, B. (2009). Examples for RDA – compared to AACR2 (work in progress). For Atlantic Provinces Library Conference. http://www.louduggan.ca/apla2009.ca/images/presentations/rda_examples.doc.

VIAF. (2010). Virtual International Authority File. http://www.viaf.org/.

Zeng, M. L., & Chan, L. M.. (2006). Metadata interoperability and standardization -- a study of methodology part II: achieving interoperability at the record and repository levels. *DLIB Magazine*, Volume 12, Number 6. http://www.dlib.org/dlib/june06/zeng/06zeng.html.