A Crosswalk from ONIX Version 3.0 for Books to MARC 21

Carol Jean Godby

Senior Research Scientist
OCLC Research

A publication of OCLC Research

A Crosswalk from ONIX Version 3.0 for Books to MARC 21
Carol Jean Godby, for OCLC Research

© 2012 OCLC Online Computer Library Center, Inc.
Reuse of this document is permitted as long as it is consistent with the terms of the Creative Commons Attribution-Noncommercial-Share Alike 3.0 (USA) license (CC-BY-NC-SA): http://creativecommons.org/licenses/by-nc-sa/3.0/.

May 2012

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

ISBN: 1-55653-434-5 (978-1-55653-434-8)
OCLC (WorldCat): 747281804

The related crosswalk is issued under a Creative Commons Attribution 3.0 (CC-BY) license: http://creativecommons.org/licenses/by/3.0/.

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

Suggested citation:
Godby, Carol Jean. 2012. *A Crosswalk from ONIX Version 3.0 for Books to MARC 21*. Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2012/2012-04.pdf (report) and http://www.oclc.org/research/publications/library/2012/2012-04a.xls (crosswalk).

Contents

Tables

The following tables are available in the ONIX 3.0 for Books–MARC 21 Mapping (Crosswalk) spreadsheet, available online at http://www.oclc.org/research/publications/library/2012/2012-04a.xls.

ONIX 3.0
ProductForm
ProductContentType
ProductFormDetail
ContributorRole
Edition Type
Extent
BISAC SH
Audience Code
Audience Range
Description
Measure
Country Codes
State Provinces
RelatedProduct
Subject
Language

Figures

Executive Summary

This report presents the crosswalk developed at OCLC for mapping the bibliographic elements defined in Version 3.0 of ONIX for Books to MARC 21 with AACR2 encoding. It is an update to the previous report Mapping ONIX to MARC, which was published in 2010 and focused on ONIX 2.1. The new version of ONIX requires a major revision to the crosswalk because of the following changes introduced in ONIX 3.0: the removal of deprecated elements supported on ONIX 2.1, a newly defined syntax that is not backwardly compatible with the ONIX predecessors, expanded support for e-books and marketing collateral, a newly defined and slightly simplified representation of collections, and many minor changes that support more detailed and internally more consistent descriptions.

The introductory sections describe the layout of the crosswalk and a strategy for deriving ONIX 3.0 syntax from ONIX 2.1, concluding with a note about how the translation logic is implemented at OCLC. The core section builds a production-quality MARC record from ONIX 3.0 input, focusing on relationships that constitute an important source of shared value in the library and publisher communities, are newly introduced or extensively revised in ONIX 3.0, or illustrate unresolved conceptual problems with bibliographic description. In the process of creating this record, it will become apparent that some of the concepts introduced in ONIX 3.0 are not easily expressed in a MARC record with AACR2 semantics. The final section speculates on how the MARC output might be improved if the AACR2 semantics is replaced by RDA.

Despite these changes, however, the MARC record that is produced from the new crosswalk is only slightly different from the output generated by the previous version. This is because the semantic impact of the changes introduced by ONIX 3.0 to elements that support a bibliographic description are relatively slight, and are, in some cases, closer to the semantics of the MARC record. Nevertheless, it is impossible to lose sight of the fact that this project was undertaken at a time of tremendous flux in the evolution of standards bibliographic description as the library community seeks to implement new content standards such as Resource Description and Access, and as standards experts consider the arguments for representing bibliographic descriptions as linked data to support better models of information exchange among libraries and the publisher supply chain.. The mappings do not reflect these changes, but the work described in this report needs to be done to create the foundation for this new paradigm. And once it is more solid, the work that remains can be carried out by a much less complex solution than the mappings described here.

1.0 Introduction

This article describes OCLC's experience with mapping ONIX 3.0 for Books to MARC and updates the 2010 report *Mapping ONIX to MARC* (Godby 2010b), which focused on ONIX 2.1. As in the earlier work, the goal is to define and make public a specification for creating a production-grade bibliographic record appropriate for the library community from publisher-supplied metadata. These specifications are represented as a table of maps between pairs of elements or fields in the two standards, or a *crosswalk*, which asserts that ONIX <Subject> is roughly equivalent to MARC 650 $a, ONIX <Contributor> to MARC 100, and so on. When supplied to a software process, these maps form a set of translation rules that convert ONIX input to MARC output. For readers who are unfamiliar with crosswalks, the Wikipedia article *Crosswalk (metadata)* gives a clear description of the concept and the surrounding issues.

The latest version of ONIX was first introduced by EDiTEUR in 2009 (EDItEUR 2009a) and is gaining critical mass as the metadata standard of choice for descriptions that support the tracking of books and related objects in the publisher supply chain. According to EDItEUR, ONIX 3.0 represents a major upgrade over Version 2.1. In technical terms, this means that version 3.0 is not backwardly compatible with earlier versions . But in conceptual terms, this seemingly abrupt change implies both that the marketplace for published products has evolved significantly and that computer systems for tracking them have become more sophisticated. As the EDItEUR principals Mark Bide and David Martin said in a webinar sponsored by the Book Industry Study Group (BISG 2009), the current marketplace has more digital products, more products that belong to a series or a set, and much more marketing collateral such as web resources, license agreements, and promotional materials in a variety of formats that are related to one another and the primary object of interest in complex ways. As a result, the sections of the ONIX standard that describe these materials have been extensively revised.

At the same time, ONIX 3.0 introduces design changes that make it easy to apply updates to existing records and to create more detailed as well as more internally consistent descriptions. To accomplish the first objective, the elements that comprise an ONIX 3.0 <Product> are organized into separate blocks that can be transmitted as separate updates. For example, the <DescriptiveDetail> block contains most of the elements of bibliographic description that have

relatively straightforward correspondences in a MARC record. <CollateralDetail> contains links to summaries, reviews, biographical statements, tables of contents, and other resources that give users a glimpse into the product's intellectual content. And <SupplyDetail> contains data indicating the availability of the primary object in the marketplace. The second objective is achieved by permitting some critical ONIX elements such as <PersonName> or even the structured name components such as <PersonNameInverted> to appear in multiple contexts. The elements required to associate a name with an identity such as first name, last name, title, address, affiliation, date of birth, and so on, were originally used to define contributors for the primary object. But <PersonName> can now be used to describe a contributor for a member of a series or collection, an author of a review, or even a personal-name subject such as *Johann Sebastian Bach*. In earlier versions of ONIX, only the first context for a personal name was fully specified using the full range of elements for describing personal names. The revamped design rpresents what Graham Bell, chief data architect at EDItEUR, calls *modular entitity descriptions*, which can be modeled more naturally in entity-relationship models and positions ONIX for a future representation as RDF or linked data. This issue will mentioned again briefly in the next section.

1.1 A major upgrade and an aging standard

Despite major changes in the ONIX standard, however, the MARC record produced from the updated crosswalk is only slightly more detailed than the results obtained from OCLC's previously defined mapping from ONIX 2.1. But the reasons underlying this conclusion are complex and not immediately obvious.

It is instructive to anchor the discussion to concepts defined in the Functional Requirements for Bibliographic Records, or FRBR, the theoretical framework developed under the auspices of the International Federation of Library Associations (IFLA) that highlights similarities in content across a myriad of physical differences. In the FRBR Group I framework (IFLA 2011), four entities represent intellectual or artistic output at various levels of abstraction. An *item* is the most concrete; it is the resource that is bought, sold, borrowed, read, watched, or listened to. Next is a *manifestation*, or a particular edition or production run that produces a set of identical items to which a standard identifier is usually assigned. Still higher is an *expression*, such as a tranlation, adaptation, or performance. Together, these concepts enable librarians to talk about the huge variety of ways that patrons can experience a work such as *Hamlet:* as a written text translated into hundreds of languages, as a synopsis or critique of the text, as a live performance of either the entire text or an excerpt, or as a sound or video recording of one of these performances.

1.1.1 ONIX Products and FRBR manifestations

Though the descriptions in ONIX and MARC are, in many respects, complementary, the two standards share the goal of describing the concepts defined in a FRBR manifestation. This relationship is illustrated schematically by Bell (2012) and reproduced in Figure 1.1 below. For the sake of completeness, this diagram also refers to the INDECS model, which was developed at the same time as FRBR and shares many of the same high-level concepts (Rust and Bide, 2000). At the top of the two models, the distinctions among FRBR works and expressions are represented equivalently in INDECS as expressions that may be related to other expressions by values such as *translation-of*. To represent information at this level, ONIX maintains a format for registering an International Standard Text Code, or ISTC (EDItEUR 2009d), which is a relatively new identifier for labeling the unique textual content of a work independently of its physical representation as a printed volume, file format, or other container. At levels of abstraction below the manifestation, ONIX and MARC differ because the metadata must reflect the fact that the libraries and publishers have different priorities in managing, locating, and delivering items.

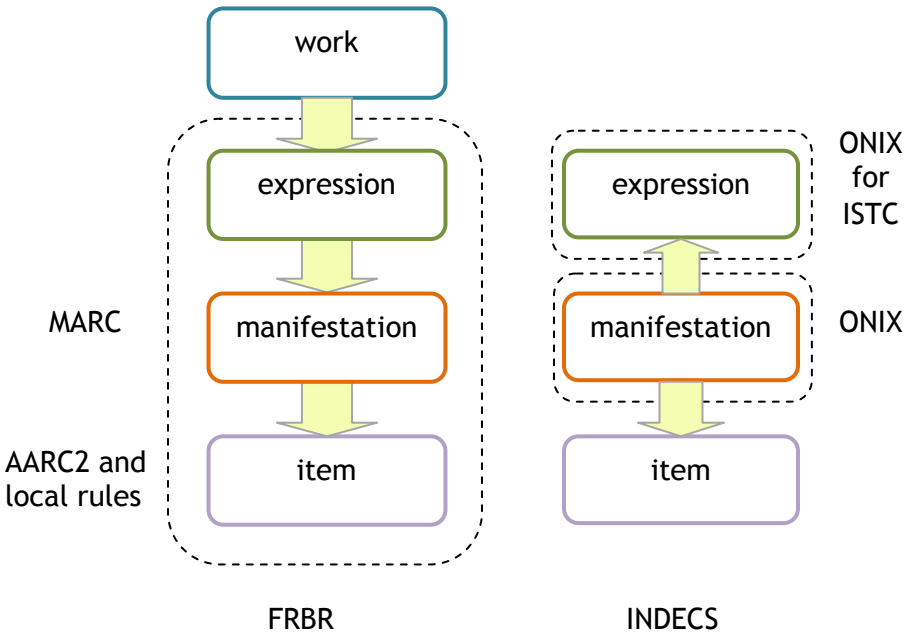

Figure 1.1 MARC, ONIX, and the FRBR hierarchy

Thus the simplest explanation for the conundrum that the new mappings produce only marginally different MARC records, despite the fact ONIX 3.0 represents a major upgrade over ONIX 2.1, is that the bibliographic description of the manifestation is largely unchanged. In

fact, except for the addition of more marketing collateral, as described in Section 3.5 below, the ONIX 3.0 bibliographic description has moved closer to the library standard. For example, Section 3.1 describes the newly defined <Collection> composite, which can align more closely with the MARC Series concept, but can also describe a group of objects bundled for sale that has no meaning in a library context. As another example, the ONIX 3.0 treatment of personal names now has links to library authority files and emerging standards such as the International Standard Name Identifier, or ISNI (2009), as well as to structured names in contexts other than the <contributor> composite, such as subject headings. These changes represent improvements that will ease the task of identifying and tracking named entities as they pass through systems maintained by the library community and the publisher supply chain.

1.1.2 Unsettled practice: digital products

Nevertheless, one of the most important innovations in bibliographic description introduced by the new version of ONIX is not reflected in the mapping to MARC. Responding to cultural and technological changes that affect both publishers and libraries, ONIX 3.0 permits much more detailed descriptions of digital products, but relatively few of the newly defined elements have a place in a MARC bibliographic record, as Section 3.4 argues. There are two reasons for this discrepancy. First, many of the new elements describe availability and usage constraints that fall outside the scope of a bibliographic description. But more significantly, the library and publishing metadata communities express genuine uncertainty about how best to describe these materials. If the digital product is a book, the concept may not have a clear referent in the minds of customers or library patrons. How is a book different from an article or other document? Is the digital product a book if it consists mostly of pictures or interactive multimedia? As a result of these uncertainties, the ONIX 3.0 best-practices guide for the description of digital products (EDItEUR 2009c) defines e-books by a set of features: the format is *online* and *digital* and the primary content is *text*. Readers interested in the size or length of the e-book might encounter a few ONIX or MARC descriptions that refer to the page count of the print antecedent, but this practice is not established.

More unresolved issues revolve around the need to reconcile the mismatch that occurs when identical content is represented in different file formats required for incompatible devices such as Kindles, Nooks, and iPads. Does each format require a different ISBN? If so, does each ISBN trigger a new metadata record? Does this sequence of events cause problems for a library, which aims to facilitate access by listing multiple URLs in a single record that describes the content? Does this practice, in turn, break the correspondence between a bibliographic record and a manifestation? Since all of these questions are unresolved, the relationship between the treatment of digital products in ONIX and MARC described in this

document is conservative and even minimal, representing a snapshot of consensus that will certainly change as metadata best practices stabilize.

1.1.3 Mapping a new version of a standard to a retiring one

Perhaps the most obvious point in considering why the enlarged ONIX standard does not produce more detailed MARC records is that the task of mapping ONIX 3.0 to MARC is essentially an exercise in mapping the newest version of the input standard to an aging target. In this view, MARC cannot express the newly defined ONIX concepts even if they fall within the scope of a bibliographic description. In the past two or three years, the library metadata standards community has moved rapidly toward the consensus that MARC is near the end of its productive life (LC 2011). Nevertheless, many experts estimate that MARC will be around for least ten more years, despite the fact that new content standards such as Resource Description and Access, or RDA (JSC-RDA 2009 and RDA 2011) have already been proposed, in part with the goal of easing the task of sharing library metadata with publishers and other communities.

For the foreseeable future, then, libraries will need to merge library and publisher metadata using mappings like those proposed in this document. Though this effort serves the pragmatic need to maintain a legacy standard, it can be planned with the future in mind. For example, the consensus implied in a set of mappings designed to minimize loss of information can influence the development of both standards. Thus it is not inconceivable that future versions of ONIX could be enhanced with more placeholders for authority-controlled elements or the FRBR-inspired ontologies that specify a broader range of expressions and manifestations than can be expressed now. On the other hand, it is worth working out the equivalences between combinations of MARC fields and the more transparently defined ONIX elements, not only because of the pressing tasks at hand but because many of these relationships are reversible, affording us an advance peek at how library metadata might be expressed in a more modern format. To prepare for this future, Karen Coyle (2012) argues that a post-MARC descriptive standard should be easier to process algorithmically, using machine-readable lists and schemas as well as a greater reliance on data, not textual descriptions. It should have identifiers, not text strings for ISBNs and other alphanumeric tokens. And descriptions conforming to this standard should consist of a set of individual statements about titles, authors, subjects, and so on, that can be easily remixed—and no longer assume the form of a record with large numbers of interdependencies that is more difficult to dismantle and reassemble. Since the ONIX design already has these features, an ONIX view of MARC data can set an example, or at least be interpreted as evidence that the post-MARC bibliographic description need not be invented from scratch. And since linked data is recognized as the ultimate destination of this effort by metadata standards experts in many industries, there should eventually be greater consensus about how to express bibliographic metadata, raising

hope that the work that remains will be carried out by a much less complex solution than the mappings described here.

Because so many issues remain unsettled, this report has several goals: to provide an interpretation of the crosswalk that may not be easily obtained by reading lines in a table; to enable metadata experts in the library and publishing communities to assess the promise as well as the limitations of mapping between the two most sophisticated standards for describing intellectual output; and to enable experts in the standards communities to identify places where the standards might be improved. But the curious layman might also find something of interest here because this is the study of how to describe the objects that touch the lives of anyone who looks for a book in the library, places an order on a bookseller website for a Blu-Ray DVD, or wishes to find out if the forty-year-old paperback novel he has just unearthed in a used bookstore has ever been made into a movie.

1.2 A note about organization

This report has four sections. Section 2 describes the layout of the crosswalk and a strategy for deriving ONIX 3.0 syntax from ONIX 2.1, concluding with a note about how the translation logic is implemented at OCLC. Section 3 builds a complex MARC record from ONIX 3.0 input, focusing on relationships that constitute an important source of shared value in the library and publisher communities, are newly introduced or extensively revised in ONIX 3.0, or illustrate unresolved conceptual problems with bibliographic description. In the process of creating this record, it will become apparent that some of the concepts introduced in ONIX 3.0 are not easily expressed in a MARC record with AACR2 semantics. Section 4 speculates on how the MARC output might be improved if the AACR2 semantics is replaced by RDA.

2.0 Design and Implementation of the Crosswalk

The translation from ONIX 3.0 source elements to MARC targets retains the scope and coverage of the earlier translation from ONIX 2.1 and extends it with as many of the newly introduced descriptors that MARC can support. The complete crosswalk is available from the spreadsheet that accompanies this document. Many of the individual associations, or *maps*, are one-to-one correspondences between ONIX and MARC modes of expression. But given that ONIX relies heavily on controlled codes that enable a human reader or software process to interpret the data, most of the maps involve some conditional logic. All mentions of ONIX codes in the text below refer to Codelist Issue 12 (EDItEUR 2010). Conversely, all descriptions of MARC record specifications are derived from *MARC 21 Format for Bibliographic Data*, published by the Library of Congress (LC 1999).

2.1 Design

The spreadsheet has 17 worksheets. The worksheet named ONIX 3.0 gives a high-level view of the crosswalk, while the others have tables or detailed algorithms for generating MARC fields or data values. In the ONIX 3.0 worksheet, three colored bars indicate the major divisions in the ONIX record and permit the reader to guage the scope of the crosswalk to MARC. The green horizontal bands mark large divisions such as <Header>, <Product>, and the subdivisions of <Product> introduced in ONIX 3.0: <DescriptiveDetail>, <CollateralDetail>, <PublishingDetail>, <RelatedMaterial>, and <SupplyDetail>. The yellow bands mark the major components of a bibliographic description such as product numbers, product forms, titles, collections, authorship, subjects, and so on.

The grey blocks in the spreadsheet represent the ONIX elements that have not been mapped. In most cases, these elements are outside the scope of a librarian's bibliographic record because they contain details about availability or commercial transactions. Or they specify access restrictions that are typically negotiated by license agreements that cover bulk orders, not individual items. In other cases, the ONIX elements contain potentially relevant library data, but they contain more detail than the current version of MARC with AACR2 semantics can support. For example, the name identifier composite includes a reference to the International Standard Name Identifier, or ISNI, a standard that is being considered for adoption in library metadata. And the detail available from ONIX 3.0 records about titles,

authors, and physical formats of review and other collateral text in the <TextContent> composite would be relevant only in a library database that is richly populated with user-contributed text.

A closer look at the worksheet named *ONIX 3.0* reveals that it is organized into five columns. Column A divides the worksheet into major blocks and is included only for readability. Columns B-G give details about the ONIX source elements. Columns B-E provide a complete list of elements in the eye-readable long format, presented in the order required by the ONIX 3.0 XML schema; Column F annotates the meaning of each element and, where relevant, the codelist that populates it; and Column G indicates whether the element is required or optional. Column H specifies the map to MARC. In most cases, the map requires simple conditional logic because ONIX relies heavily on coded values to specify the interpretation of a data element. For example, consider the relationship between descriptions of standard identifiers in lines 40-43. Here the crosswalk specifies that the MARC field generated by <Productidentifier> depends on the <IDValue> code. If it is 15, indicating that the identifier is ISBN-13, the MARC field 020 $a is generated; if it is 06, indicating a digital object identifier or DOI (DOI 2011), the MARC field *024 7 $a {doi-value} $ doi* is created; and so on, for all of the correspondences listed in Column H of row 42.

But when the logic is too complex to be readable in a compact tabular format, Column H points to the worksheet that contains a detailed algorithm. For example, the map from <RelatedMaterial><RelatedProduct> (line 424) points to the worksheet named 'RelatedProduct' that describes which MARC linking field to generate and which MARC subfields it should contain. And the worksheet named 'ProductForm' describes a map from the ONIX <ProductForm> to MARC values in the Leader, 007, 008, 245, and 300 fields. These maps are also discussed in Sections 3.3 and 3.5 below.

2.2 Implementation

As described in an earlier publication (Godby, Smith, and Childress 2008), OCLC's implementation defines an XML scripting language called *the Semantic Equivalence Expression Language*, or Seel, which represents the information in a map defined in a crosswalk and submits it for execution to a locally designed interpeter with the same name. The smallest executable script is a single <map> element, which contains a source, a target, and an optional set of conditions. The largest executable script is an arbitrarily large set of maps collected under a <translation> element. With this design, it was a relatively straightforward editing task to modify maps containing ONIX 2.1 paths to create ONIX 3.0's richer container structure, add the newly defined elements, delete obsolete elements, rename elements, or reposition them in the output record.

Figure 2.1 shows how a Seel script for part of the ONIX 2.1 <Title> Composite is converted to the corresponding ONIX 3.0 <TitleElement> composite by adding extra depth to the source path. As in the ONIX 2.1, a *MARC 245 $a* field is generated only if the ONIX <TitleType> is 0, a condition recorded in the Seel <context> element. The examples described in the next section of this article are much more subtle and the 2.1 and 3.0 structures are less obviously related, but the essential operations remain the same.

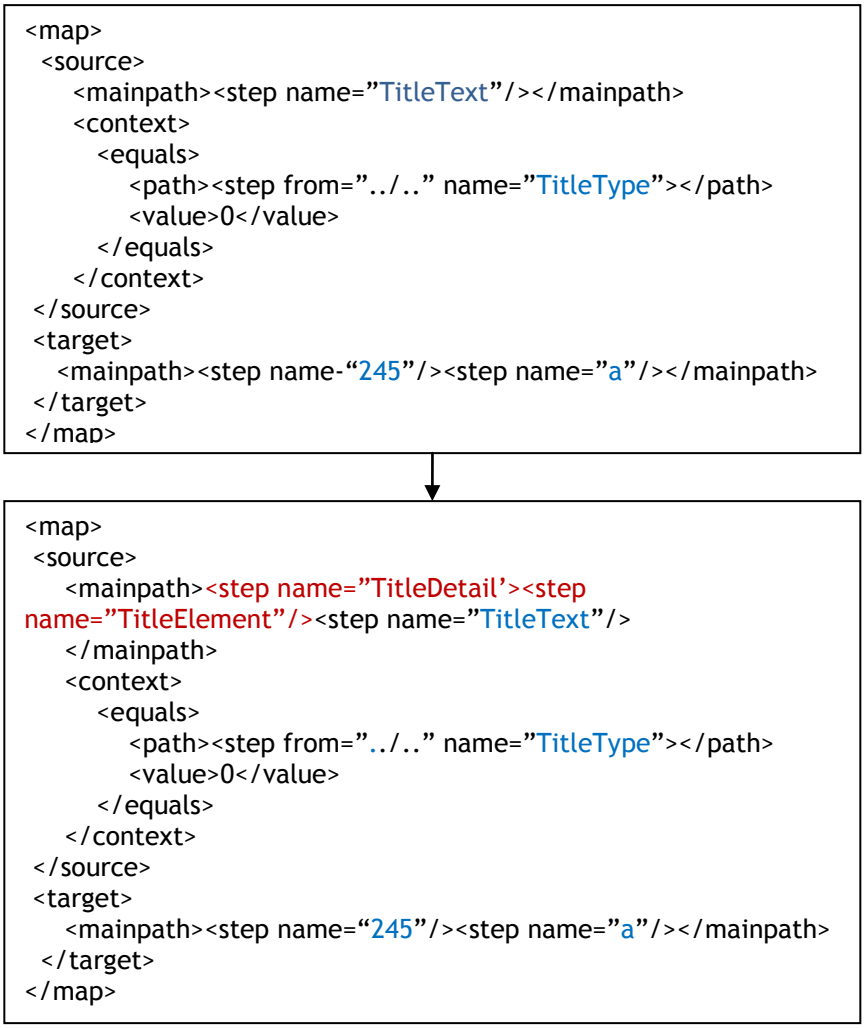

Figure 2.1 A Seel script for mapping a primary title

To make the task of defining maps from ONIX 3.0 manageable and to promote consistency with the ONIX 2.1 translation, we observed two guidelines. First, we strove to reuse as much of the logic from the ONIX 2.1 translation as possible. We were able to retain nearly all of the previously defined MARC outputs and concentrated our effort on regenerating them using ONIX 3.0 input. As Figure 2.1 implies, the <target> elements in the maps are mostly

unchanged but the <source> elements have been modified. Second, we added maps for newly defined elements in ONIX 3.0 only if they could be placed into highly structured and easily parsed MARC fields. By following this guideline, we refrained from defining new maps to the MARC 500 (General Note) field, which is already overloaded (Smith-Yoshimura et al. 2010) and to the 887 (Non-MARC Information) and 900 (Locally defined) fields, whose meaning differs from one usage or installation to the next.

3.0 Creating a MARC Record from ONIX 3.0

The extended discussion in this section singles out for special comment the maps to MARC from the ONIX 3.0 elements that are new, have been extensively revised, are important for sharing information between the library and publishing communities, or are complex and problematic. The expository strategy is to present a series of vignettes, which can be read in isolation as case studies that might be further developed with advances in best-practices guidelines. Taken together, they form the essential components of a description of the book *The Laboratory Canine*, written by Garett Field and Todd A. Jackson and published by CRC/Taylor & Francis in 2006. This book is a member of the *The Laboratory Animal Pocket Reference Series* and is now available as an e-book in the Amazon Kindle format. The final segment completes the MARC record with elements that can generated programmatically from the editing steps describe above and submitted to the same algorithms we developed for the ONIX 2.1 version of the crosswalk.

3.1 Series and collections

The discussion begins with the treatment of series and collections because this topic raises a fundamental issue that pervades the rest of the description: is the description about a single product, or about a group of products that are related by some well-defined criterion? In a MARC record, this decision is recorded in the value of the '*Bibliographic Level'* byte of the *Leader* field, which specifies whether the object being described is a single item, a member of a collection, or a subdivision of a larger work, such as a chapter in a book. Essentially the same distinction can be made in ONIX, though the goal is simply to indicate membership in a collection and not to describe it, as the corresponding MARC record might.

Figure 3.1 shows the effect of the maps from ONIX <Collection> on an ONIX input and a corresponding MARC output record describing the book *The Laboratory Canine*. The maps start at line 106 in the tab labeled *ONIX* in the crosswalk.This record is based loosely on an ONIX 3.0 record processed at OCLC from a file created by Taylor and Francis.

To promote readability, Figure 3.1 also illlustrates some conventions used throughout Section 3 of this document. First, where it is relevant to the discussion, ONIX 2.1 source is shown for comparison. Second, the figures and the explanatory text refer to individual *elements* or

fields, such as ONIX <TitleText> and MARC 245; or to an ONIX *composite* such as <Collection>, which is a hierarchically arranged group of elements that describe components of single concept such as *collection*, *title*, or *subject*. To simplify the examples, the major ONIX blocks <DescriptiveDetail>, <CollateralDetail>, and <RelatedWork> are omitted in record fragments and the focused discussion, but they appear in the complete record shown in Figure 4.2. Finally, the text colors are meaningful and follow the same conventions I defined in the description of maps from ONIX 2.1 (Godby 2010b): the blue text signifies structural features of the two standards; the red text is the literal data that is copied from the source to the target records; and the green text represents machine-processable codes that are used to interpret the narrow meaning or format of an element and are often used to state a condition on a map.

The right side of Figure 3.1 shows the ONIX 3.0 input. For comparison, the left side shows the corresponding representation as a <Series> composite in ONIX 2.1. Note that the ONIX 3.0 <TitleElement> composite has a <TitleElementLevel>, where the value of '2' indicates that *The Laboratory Animal Pocket Reference Series* is a title of a series, while the value of '1' for *The Laboratory Canine* is identifies this string as the title of an individual product. Both ONIX sources generate two fields in a MARC record: the 245 $a field for the Product title and the 490 $a field for the Series or Collection title. The value 'a' in the MARC Leader field indicates that the primary product described in the record is an item or monograph, which can be inferred from the <TitleType> value of '1' for the topmost title in the ONIX description.

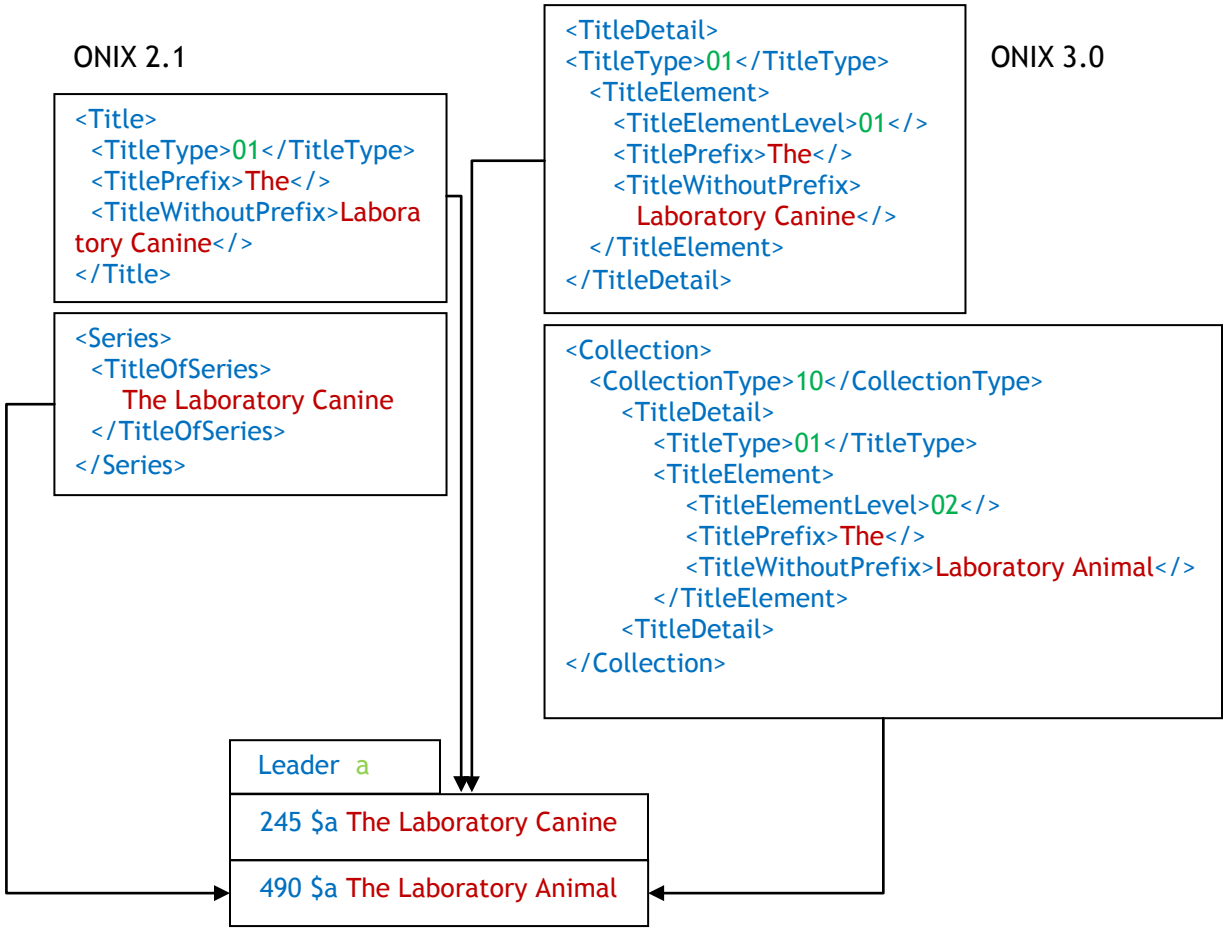

Figure 3.1 Maps from collections and series in ONIX 2.1 and 3.0

But this map masks a major revision from ONIX 2.1 to 3.0. The concept of *Series* appears to be shared between libraries and publishers because it is commonly understood as a group of products that are published over an indefinite time period and grouped together under a series title for marketing purposes, and not traded as a single item, such as the books that comprise *Nancy Drew Mystery Stories* published between 1930 and 2003 by Grosset and Dunlap. The <Series> composite was revamped as <Collection> to subsume commonalities between sets and series or ad-hoc associations between products. This change is potentially confusing to the library community because it obscures the once shared concept of Series and includes many associated products that imply no curatorial intention.

Fortunately, a series can still be identified as a type of collection. According to the best-practices guideline for describing sets and collections (EDItEUR 2009b), this is accomplished with the required element <CollectionType>. The OCLC crosswalk maps the collection to a MARC Series statement only if the <CollectionType> value is 10, which specifies a *publisher collection* and is defined as "a bibliographic collection to which the publisher assigns a

collective identity," such as Penguin Classics (Penguin Group 2011). A *bibliographic collection*, in turn, is one in which each member has a bibliographic description. Publisher collections are also distinguished from *ascribed collections*, which are bibliographic collections created by an entity other than a publisher, typically a metadata aggregator.

With this detail established, the rest of the map from the <Collection> composite is straightforward and semantically closer to the MARC Series statements than the original map from the ONIX 2.1 <Series> composite. As Figure 3.1 shows, the translation to the MARC Series statements remains the same because only the ONIX source has changed and the differences can be derived programmatically, despite major differences in appearance between the ONIX 2.1 and 3.0 composites. Thus <Series> is equivalent to <DescriptiveDetail><Collection> and <TitleOfSeries> is equivalent to <Collection><TitleDetail>, and so on. The destination is still a MARC record with an 022 field containing the ISSN and a 490 field that has placeholders for titles, identifiers, and volume numbers.

Figure 3.1 also illustrates a major improvement to the structure of ONIX 3.0, which is perhaps motivated by the commonsense observation that the most important elements of a bibliographic description such as titles, personal names, subjects, and product forms have a constancy of meaning regardless of where they are used in a record. These elements, reflecting the improvements imposed by the introduction of so-called *modular entity descriptions* mentioned near the end of Section 1.0, are now matched by a constancy of form. In Figure 3.1, the titles of the primary product and the collection can be described using the same rich <TitleDetail> composite instead of separately defined <Title> and <TitleOfSeries> elements. Nevertheless, the MARC 490 $a destination cannot capture the additional detail because it is an undifferentiated string.

Though more work remains to be done to realize the full promise of modular entity descriptions in ONIX 3.0, they represent an advance toward the goal of defining an interoperable vocabulary that is independent of the surrounding context and bears some resemblance to the design of Dublin Core Terms (DCMI 2011). The words in this vocabulary behave much like words in a natural language, which are defined as combinations of symbols whose meaning and form remain constant and are the building blocks from which larger expressions are built.

3.2 Subjects

Subject headings and classification codes represent an important source of shared value between the library and publisher communities. The mappings of the Subject elements are complex but derived programmatically and suffer from relatively little loss of information. They also form an infrastructure for even more added value, such as the mappings between

Dewey and BISAC codes maintained by the editorial staff of the Dewey Decimal Classification (Mitchell 2010). As a result of the changes introduced by ONIX 3.0, <Subject> composites support richer descriptions and are more efficiently processed.

Figure 3.2 shows a map to MARC from an ONIX primary subject coded in versions 2.1 and 3.0. A BISAC subject code is mapped to a MARC 072 field, preserving the algorithm devised for the ONIX 2.1 version of the map. As before, the BISAC code is interpreted as a subject classification with two levels of hierarchy and represented in the MARC 072 field, the *Subject Category Code*. In this example, which continues the description of *The Laboratory Canine*, the alphabetic prefix MED identifies the broad category *Medicine* (BISG 2011), while the numeric code 089000 representing the subcategory *Medical/Veterinary Medicine/General* and is listed in $x, the subfield for a subject category code subdivision. The corresponding text is copied in a 650 $a field, which is tagged by the second indicator value of *7* and the $2 value of *bisacsh* as a BISAC subject heading. In other cases, BISAC codes contain information that can be used to populate the MARC 008 fields. For example, the BISAC prefix JUV sets a 008 *Target audience* value for juveniles, while BIO sets a value for *Biography* indicating the presence of biographical information. The 2011 revision of the BISG codes shows a move toward a three-level hierarchy, but this change is not completely implemented and not reflected in the mappings described here.

In the ONIX 3.0 representation, the <Subject> composite is used for the main subject instead of the dedicated 2.1 element <BASICMainSubject> and is set apart from other <Subject> composites by the empty element <MainSubject/>. Though the implementation of the format change requires only a mechanical conversion, it masks a subtle semantic and pragmatic difference between the two representations. Though technically not enforced by the ONIX 2.1 XML schema or DTD, the <BASICMainSubject> is nevertheless a best-practice recommendation for conformance with the BISAC content standard, which must be populated with a BISAC code; a corresponding assignment of BIC headings is required in the United Kingdom (BIC 2011). In the ONIX 3.0 version of the record shown in Figure 3.2, the SubjectSchemeIdentifier of 10 still tags this code as BISAC, but BISAC and BIC no longer have a privileged status. Now a broad range of subject headings schemes is admissible as a <MainSubject> and it is up to the metadata providers to supply the type of heading expected in their communities. This broader scope is a closer match to the semantics of the MARC 072 field.

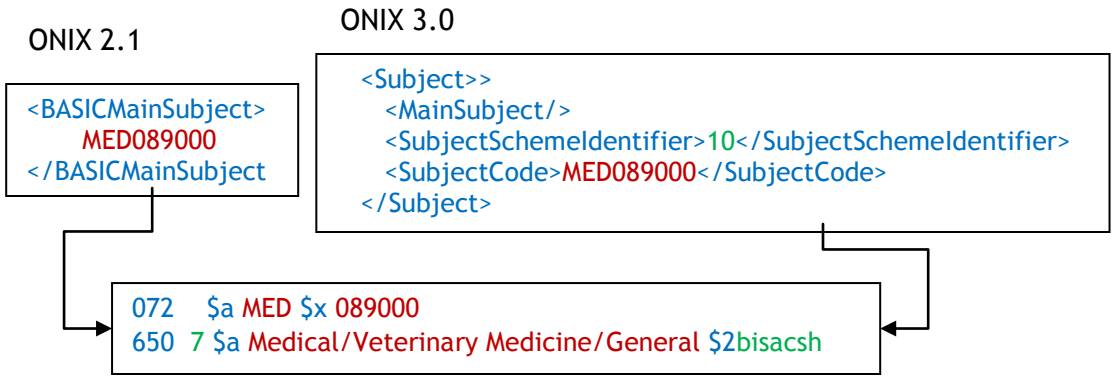

Figure 3.2 An ONIX Main subject mapped to a MARC classification code

To extend the discussion of subjects, Figure 3.3 shows the effects of the map from the MARC 600 field and highlights the differences in the treatment of personal names in ONIX 2.1 and 3.0. The main issue is that ONIX 2.1 defines a large set of elements for describing the personal name in the <Contributor> composite, though only an undifferentiated string is available for the <PersonAsSubject>. In ONIX 3.0, this discrepancy is corrected by recasting the <NameType> elements as modular entities, here exemplified by the <PersonNameInverted> element. Both standards now support equally rich descriptions of personal names as contributors or subjects. In the subject map shown in Figure 3.3, the personal name string in <PersonNameInverted> is mapped to MARC 600$a, while the birth and death dates are mapped to the MARC 600 $d field. Here the ONIX 3.0 description matches the granularity of the corresponding MARC description.

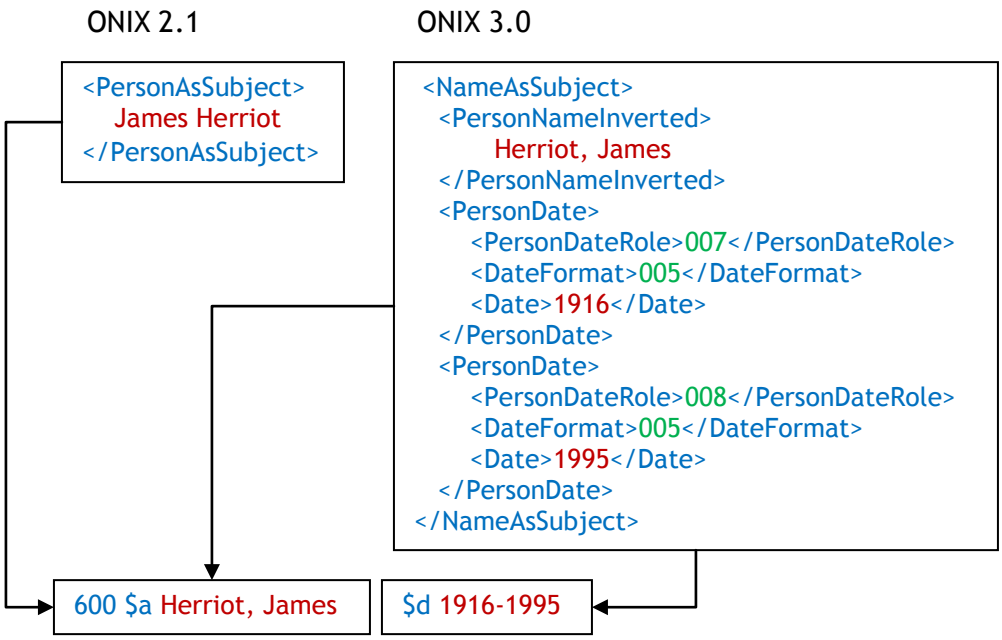

Figure 3.3 A personal name subject

Even more detail from ONIX 3.0 can be mapped because both standards now have elements for titles and honorifics, affiliations, and authority record control numbers. But ONIX is slightly more flexible because it labels the subparts of a name separately, while the corresponding slot in the MARC standard is the undifferentiated string in $a. As a result, names that present special problems for indexing or alphabetizing, such as *James van Buren, Jr., Ph.D.*, are assembled from an ordered sequence of the ONIX elements <KeyName>, <SuffixToKey>, <LettersAfterNames>, <NamesBeforeKey>, and <PrefixToKey>, producing the MARC $a value *Buren Jr., PhD., James van*.

3.3 Product forms

The <ProductForm> element receives the same treatment as in the ONIX 2.1 crosswalk and is enhanced with the expanded coverage of e-books and the <DescriptiveDetail> container required by ONIX 3.0. As in ONIX 2.1, the mapping is one-to-many: a single <ProductForm> value obtained from a list of formats specified in Codelist 150 triggers the assignment of as many as twenty data values in the MARC Leader, 007, 008, and 300 fields. Thus a core problem persists that has been described in earlier accounts: establishing a reference (Godby 2010a), which is especially critical for descriptions of electronically delivered multimedia. The <ProductForm> value is a selection from a list of formats that are available in the current marketplace; when mapped to MARC, this value is decomposed into a set of fourteen attributes. For example, a <ProductForm> value of AE, an 'Audio Disc (excluding CD)' according to the codelist gloss, is mapped to 007 values that specify a plastic mass-produced

disc containing a sound recording. Unfortunately, there is no certainty that these attributes are the important technical specifications that distinguish among related objects such as *records*, *45s*, *78s*, *LPs*, or *long-playing albums* or that the vocabulary used by the general public resolves either to the ONIX annotation or the MARC description. As a result, we cannot guarantee that the ProductForm or MARC 007 values be used to match a product offered for sale in the publisher supply chain to the object desired by consumers or library patrons. This problem remains unresolved, but a promising way forward is presented in Section 4.

Here I will point out another important problem, which appears to be conceptually simpler: the description of a printed book. Figure 3.4 shows the mapping to MARC. According to Codelist 150, the <ProductForm> value 'BA' identifies a 'book – detail unspecified,' according to the gloss, and is appropriate when more specific information about a book's physical format is unavailable. This is mapped to two values in the MARC Leader field, the *Type of record* (a) and *Bibliographic level* (m), which establish that the content is primarily linguistic rather than, say, musical or graphic and that the object is a single item rather than a series or collection. The 008 value *Form of item* is also relevant in this relationship and is used to distinguish printed books from e-books, a point that is elaborated in the next set of examples. But when the item is an ordinary printed book, no *Form of item* value is set because this element would be used to specify only unusual formats such as microfilm, microfiche, large print, or Braille. Unfortunately, the more descriptive Codelist 150 B-values that specify hardbacks, paperbacks, looseleaf or spiral bound notebooks, board books, bath books, big books, or books with leather or fine binding, can be captured in MARC only with an optional free-text note appended to the 020 $a field; as in, for example, 020 $a 9780060723804 (paperback); or 020 $a 0877790124 (blue pigskin). Because this information is problematic for machine processing, it is not represented in the mapping. In other words, much of the detail available from an ONIX description about the form of a printed book is lost when it is mapped to MARC.

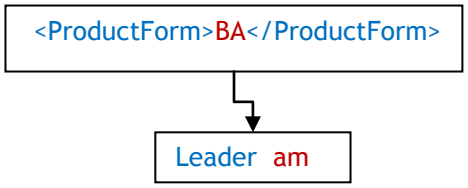

Figure 3.4 A <ProductForm> for a printed book

As in ONIX 2.1, descriptions of related products can include a <ProductForm>. Though the resulting ONIX description is parsimonious but expressive because <ProductForm> is re-cast as a modular entity, this detail cannot be mapped to MARC because the fields that contain

physical descriptors—007, 300, 008, and Leader—are available only for the description of the primary product.

3.4 E-books

Maps involving e-resources must solve three problems: distinguishing e-resources from other formats, relating an e-resource to its physical antecedent, and describing restrictions on use. Unfortunately, the third problem must simply be set aside because the newly defined ONIX 3.0 elements such as <ePubUsageConstraint> and <ePubTechnicalProtection> cannot be mapped to dedicated fields in MARC. But librarians are debating whether these elements belong in a bibliographic description. Perhaps a better place for them is a locally defined digital rights repository or a knowledge warehouse of business intelligence because they are not an intrinsic feature of the product.

Figures 3.5 and 3.6 show the results of mapping ONIX elements to MARC that solve the first two problems on a description of the Amazon Kindle version of the *The Laboratory Canine*. Though small details are lost and the MARC target is not coded as explicitly as the ONIX source, the map preserves the essential information required to locate and download the e-resource in the appropriate format. The discussion here focuses on the representation of e-books because the MARC standard has a rich set of descriptors in the 007 field for describing computer files and audiovisual materials and the issues are not the same.

Figure 3.5 shows the mapping of identifier, title, and format information. The map from <ProductIdentifier> to MARC would be unremarkable, except that MARC distinguishes between the ISBN, which appears in the 020 field; and other standard identifiers for monographs, which appear in the 024 field. Since the e-book form of *The Laboratory Canine* has no ISBN but only an Amazon Standard Identification Number, or ASIN, this value is mapped to the MARC 024 field, whose value of 8 in the first indicator labels this code as an unspecified or proprietary type, matching the designation implied by the <IDValue> of '01' implied in the corresponding ONIX <ProductIdentifier>compsite. As in earlier examples, the *01* values for <TitleType> and <TitleElementLevel> indicate that *The Laboratory Canine* is the primary title of a single product, which is mapped to MARC 245 $a without loss of meaning. The ONIX <ProductForm> value is 'ED,' indicating the product is digital content delivered only by download. The code in <ProductFormDetail> refines this description with the important detail that the format is an Amazon Kindle file, which is completed with a <ProductContentType> code designating the content as consisting primarily of eye-readable text. The use of these three codes illustrates the recommendation given in Section 3.5 of EDItEUR's best-practices guideline for the description of e-resources (EDItEUR 2009c).

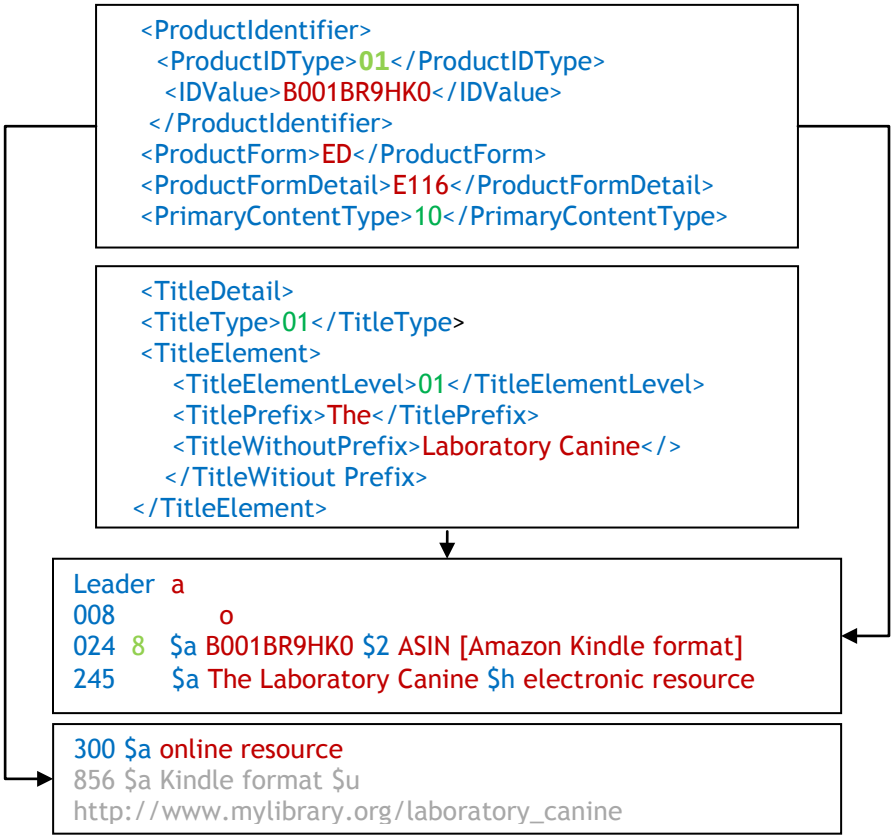

Figure 3.5 Mapping an e-book

The MARC output does not capture all of this detail because it complies with the recommendations of the *provider-neutral standard* (Martin and Mundle 2010, PNE-RTG 2009), whose goal is to minimize the number of MARC records for describing e-books. A map from all available ONIX source elements would generate a separate record for each e-book format. The provider-neutral standard eliminates a potential source of duplicated information by recommending the creation of a single record that can be used for all formats. This clash highlights a clear difference between the library patron-centered need to describe the content and the commercial need to differentiate between file formats that such as iPad and Kindle that may have different prices, require different combinations of software and hardware, and cannot both be sold by any single retailer. The MARC output shown in Figure 3.5 illustrates. The content is indicated as text by the value 'a' in the Leader field. The format is conveyed redundantly by the keyword *[electronic resource]* in 245$h; the free-text note in 300$a, *online resource*; and the 'o' value for the *Form of item* byte at position 23 in the 008, also indicating an online format. But the information in <ProductFormDetail> is not mapped. Instead, a library that has purchased rights to the Amazon Kindle version of *The Laboratory Canine* would modify the MARC record that has been produced from the crosswalk with an 856 field with two subfields: $u, a locally maintained URL from which the e-book can

be downloaded; and $q, which identifies the resource as an Amazon Kindle file. If the library later acquires rights to the Barnes and Noble Nook version, another 856 field would be added to the record. Figure 3.5 shows the 856 field in grey to indicate that is is not generated by a map from ONIX.

The result from the map for the print antecedent of the e-book is shown in Figure 3.6. In the ONIX description, this information is recorded in the segment <RelatedMaterial>, a sibling of <DescriptiveDetail>. The relationship is specified in <ProductRelationCode>, where the value of 13 indicates that the <Product> is "an epublication based on printed product <RelatedProduct>," according to the codelist gloss (EDItEUR 2010). In other words, the primary object being described, the Amazon Kindle version of *The Laboratory Canine*, is based on the print version whose ISBN-13 is tagged by <IDValue>. When this information mapped to MARC, we must accommodate the fact that there is no dedicated field or subfield for recording this relationship. The closest fit is one of the 77x linking fields, 776, which describes an 'Additional Physical Form Entry.' In this field, $z holds the ISBN, while $i contains the free-text text string *Print version:*, which is mandated by the provider neutral standard.

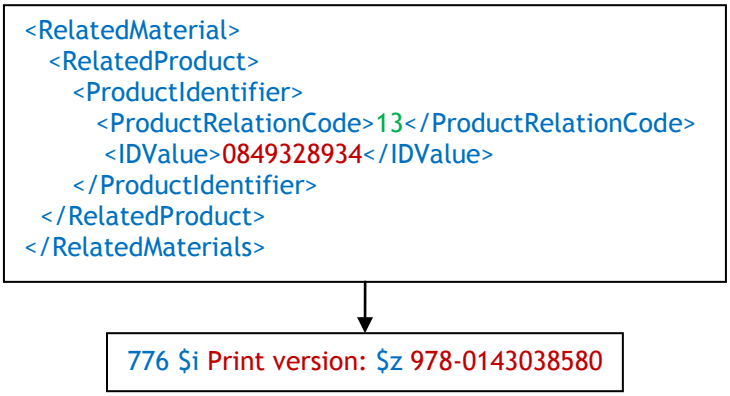

Figure 3.6 Mapping the print antecedent of an e-book

An unresolved issue in both representations is the authority of the link. In other words, is the related product the true antecedent of the e-book or simply a good enough facsimile of the content? According to the discussion generated by the webinar that introduced ONIX 3.0 to a broad audience of librarians and publishers (BISG 2009a), the curatorial work required to establish a definitive link would require a much greater investment than the marketplace demands or could support. The presence of the 767 linking field in the MARC record implies, but does not guarantee, that this work has been done because the information is not authority-controlled.

3.5 Marketing collateral

Given the ease with which the Internet, especially social media, can invite the public to evaluate and promote publishing products, an overhaul of the elements that describe the description of marketing collateral is overdue. In ONIX 2.1, marketing collateral was restricted to text that was described in the <OtherText> composite: summaries, author biographies, abstracts, tables of contents, and reviews that were generated or endorsed by the publishing community. These categories track closely with those defined for the MARC notes fields 505, Formatted Contents Note; 520, Summary; and 545, Biographical or Historical Data. In ONIX 3.0, the scope of marketing collateral has been expanded to include material in any format that can be used for promoting a product, including audio and video clips, bestseller lists, media mentions, or the full text of a review in a third-party publication. In this expansion, ONIX 3.0 renames the <OtherText> composite to <TextContent> and restricts its usage to text that is intrinsic to an ONIX record. The new material is called <CitedContent>, which is produced by a third party and accessed through a link in the record. Both composites are part of the <CollateralDetail> group.

Figure 3.7 is an example showing two pieces of marketing collateral for *The Laboratory Canine* as they would be represented in ONIX 2.1 and 3.0 and mapped to MARC: a review quote that might be generated or endorsed by a publisher's marketing staff and printed on the back cover of the paperback edition, and a second review published in the scholarly journal *Veterinary Pathology* in 2007. The material in the <CitedContent> composite is also a review, but it has a different status because it appeared after the book was published in 2006 and was created by a relevant professional with no apparent connection to the publishing industry. It is greyed out because this information is neither represented in ONIX 2.1 nor mapped to MARC.

The map to MARC is unchanged from the ONIX 2.1 version because the semantics of the corresponding MARC fields most closely match the <TextContent> composite. In Figure 3.7, the review quote is mapped to MARC 520. Note that the <TextContent> composite is isomorphic to the <OtherText> composite, from which the ONIX 3.0 map can be programmatically derived by renaming the congruent elements and enclosing them in the <CollateralDetail> segment.

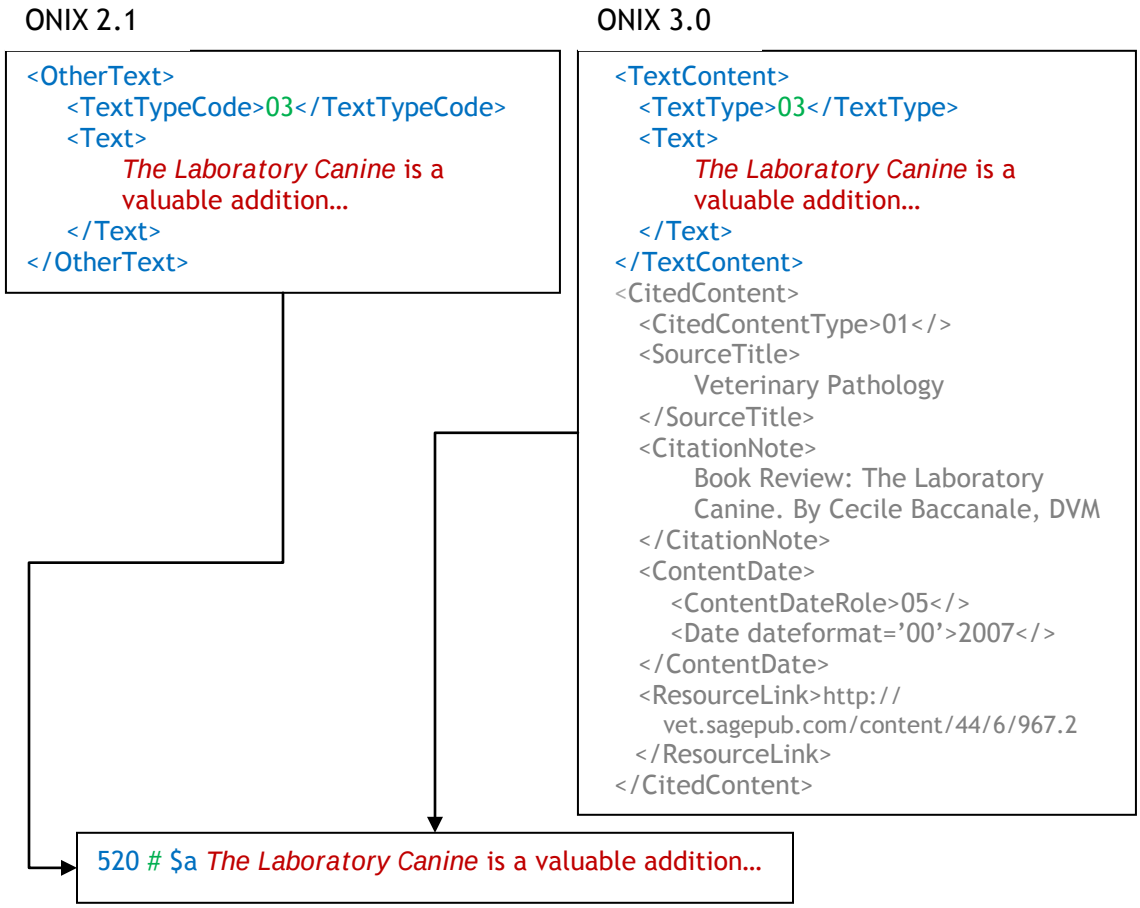

Figure 3.7 Marketing collateral that maps (and does not)

Obviously, the distinction between the two kinds of reviews is strained and it is possible to argue that the omitted review in the <CitedContent> composite is more relevant to the patrons of a research library than the review from the <TextContent> element that was mapped. But since reviews are third-party content, the differentiation reflects the need to observe copyright restrictions. <CitedContent> provides a link to the full text, while <OtherText> and <TextContent> would normally be used for small 'fair use' snippets from reviews.

Since the ONIX source is coded for machine processing, the crosswalk could be modified to populate a relevant MARC 520 field with <CitedContent> if the value of <CitedContentType> indicates a review. In that case, the link to the review could appear in the $u subfield. But the 520 $a subfield requires the text of the review, which, by definition, is not part of a <CitedContent> composite because it would be represented as a <TextContent> composite. The larger issue is that librarians and publishers both recognize that products of intellectual endeavor are best discovered and experienced in a context whose salient features are

continually being identified and formally defined in metadata standards. The <CitedContent> reviews should eventually be accessible from library interfaces because they are useful for scholarship; and, conversely, more of the curated relationships produced by librarians will appear in bookseller interfaces because they are useful for discovery. Descriptions of collateral or ancillary material promise to be a rich source of shared value between the library and publisher communities and represent one of the strongest arguments for undertaking the effort to map their standards.

4.0 Taking stock and looking forward

We can begin to assess the status of the effort to translate ONIX 3.0 to MARC by viewing a complete versions of the records that participate in the translation produced by the crosswalk that accompanies this document. The MARC target of a description of the e-book version of *The Laboratory Canine* is shown in Figure 4.1; the corresponding ONIX source is shown in Figure 4.2

```
Leader   00000 am 000008
001       0849328934
008       s2006    xx     o      000 0 eng d
024  8   $a B001BR9HK0 $2 ASIN [Amazon Kindle format]
072       $a MED $x 089000 $2 bisacsh
100       $a Field, Garrett $u Eli Lilly and Company, Indianapolis, Indiana, United States
245       $a The Laboratory Canine $h [electronic resource]
260       $b CRC $c 2006
300       $a online resource
490       $a The Laboratory Animal
521       $a Veterinary and laboratory animal technicians
520       $a The Laboratory Canine is a valuable addition to the The Laboratory Animal Pocketbook
Series.
650       $a MEDICAL/ Research $2 bisacsh
700       $a Jackson, Todd A. $u Bristol-Myers Squibb, Evansville, Indiana, USA.
776       $i Print version: $z 978-0849328930
856       $a Kindle format $u http://www.mylibrary.org/laboratory_canine
```

Figure 4.1 A complete MARC record

The maps that produce the multicolored text in both figures were described in Section 3 of this document. The elements that produce the greyed-out text are aligned by maps that have been derived from the crosswalk based on ONIX 2.1. The MARC 001, 020, and 260 fields are populated by one-to-one data transfers from the ONIX <RecordReference>, and <publisher> elements. The map from <Audience> is similar in complexity to the <TextContent> map and constructs a similar MARC note field. The maps from <Contributor> contain the most complex logic in the crosswalk but are derived from logic developed for ONIX 2.1 and are modified only slightly for ONIX 3.0. <PersonNameInverted> is the preferred ONIX source because it matches the formatting required in the MARC authorship fields, but when this element is not present, an inverted name is constructed from separately labeled elements. The first <Contributor> is

mapped to a MARC 100 field and the rest are mapped to MARC 700 fields. Despite the complexity of this relationship, little information is lost because the data models for personal names are similar in the two standards. As Figures 4.1 and 4.2 show, *The Laboratory Canine* has two authors and both have institutional affiliations. The only remaining new elements in the MARC target are values in the MARC control fields *Leader* and 008. The publication date (2006) and default language (eng) are mapped from ONIX elements, but the crosswalk also supplies defaults that identify the source and presumed trustworthiness of the record. For example, the value '8' in the Leader field represents the encoding level value for a prepublication record; and the value 'd' in the 008 field represents an unknown cataloging source, i.e., an origin that is not a cooperative cataloging program or a national bibliographic agency.

The ONIX source is shown below.

```
<Product>
    <RecordReference>0849328934</RecordReference>
    <NotificationType>02</NotificationType>
   <ProductIdentifier>
      <ProductIDType>01</ProductIDType>
      <IDValue>B001BR9HK0</IDValue>
    </ProductIdentifier>
   <DescriptiveDetail>
      <ProductComposition>00</ProductComposition>
      <ProductForm>ED</ProductForm>
      <ProductFormDetail>E116</ProductFormDetail>
      <PrimaryContentType>10</PrimaryContentType>
     <Collection>
        <CollectionType>10</CollectionType>
       <TitleDetail>
          <TitleType>01</TitleType>
         <TitleElement>
            <TitleElementLevel>02</TitleElementLevel>
            <TitlePrefix>The</Title>
             <TitleWithoutPrefix>Laboratory Animal</TitleWithoutPrefix>
          </TitleElement>
       </TitleDetail>
    </Collection>
   <TitleDetail>
      <TitleType>01</TitleType>
     <TitleElement>
        <TitleElementLevel>01</TitleElementLevel>
        <TitlePrefix>The</TitlePrefix>
         <TitleWithoutPrefix>Laboratory Canine</TitleWithoutPrefix>
      </TitleElement>
    </TitleDetail>
    <Contributor>
       <ContributorRole>A01</ContributorRole>
       <PersonNameInverted>Field, Garrett</PersonNameInverted>
      <ProfessionalAffiliation>
```

```
            <Affiliation>Eli Lilly and Company, Indianapolis, Indiana, United States</Affiliation>
          </ProfessionalAffiliation>
        </Contributor>
        <Contributor>
          <ContributorRole>A01</ContributorRole>
          <PersonNameInverted>Jackson, Todd A.</PersonNameInverted>
          <ProfessionalAffiliation>
            <Affiliation>BristolMyers Squibb, Evansville, Indiana, USA</Affiliation>
          </ProfessionalAffiliation>
        </Contributor>
        <Subject>
          <MainSubject />
          <SubjectSchemeIdentifier>10</SubjectSchemeIdentifier>
          <SubjectCode>MED106000</SubjectCode>
        </Subject>
        <AudienceCode>01</AudienceCode>
        <AudienceDescription>Veterinary and laboratory animal technicians, biomedical researchers,
          undergraduate or graduate students in biomedical sciences, and
          veterinarians</AudienceDescription>
      </DescriptiveDetail>
      <CollateralDetail>
        <TextContent>
          <TextType>03</TextType>
          <ContentAudience>00</ContentAudience>
          <Text>The Laboratory Canine is a valuable addition to the The Laboratory Animal
            Pocketbook Series.</Text>
        </TextContent>
        <CitedContent>
          <CitedContentType>01</CitedContentType>
          <SourceTitle>Veterinary Pathology</SourceTitle>
          <CitationNote>Book review: The Laboratory Canine. By Cecile Baccanale,
            DVM.</CitationNote>
          <ResourceLink>http://vet.sagepub.com/content/44/6/967.</ResourceLink>
          <ContentDate>
            <ContentDateRole>01</ContentDateRole>
            <DateFormat>00</DateFormat>
            <Date>2007</Date>
          </ContentDate>
        </CitedContent>
      </CollateralDetail>
      <PublishingDetail>
        <Publisher>
          <PublishingRole>01</PublishingRole>
          <PublisherName>CRC</PublisherName>
        </Publisher>
        <PublishingStatus>02</PublishingStatus>
        <PublishingDate>
          <PublishingDateRole>01</PublishingDateRole>
          <DateFormat>00</DateFormat>
          <Date>2006</Date>
        </PublishingDate>
      </PublishingDetail>
      <RelatedMaterial>
```

```
        <RelatedProduct>
            <ProductRelationCode>13</ProductRelationCode>
            <ProductIdentifier>
                <ProductIDType>15</ProductIDType>
                <IDValue>9780143038580</IDValue>
            </ProductIdentifier>
        </RelatedProduct>
    </RelatedMaterial>
</Product>
```

Figure 4.2 The <Product> composite in a complete ONIX record

In sum, the outcome of the crosswalk is a MARC 21 record with AACR2 semantics that can be automatically generated from an ONIX 3.0 source, pass a rigorous semantic validation, serve as a rough draft that can be further refined by cataloging best-practices guidelines, and qualify for inclusion in a quality-controlled library database. Much of the meaning of the translation from ONIX 2.1 to MARC has been preserved and is extended with concepts introduced in ONIX 3.0. The upgrade to ONIX 3.0 adds a more comprehensive treatment of e-books, a revised and simplified definition of collections, and extended definitions of marketing collateral. Though many of the elements introduced in ONIX 3.0 can be mapped to MARC, this effort raises questions about how to define the boundary of a bibliographic description because it is not co-extensive with the <DescriptiveDetail> group. Some elements that belong to other groups are also relevant to a bibliographic description, but they must be separated from the ONIX elements that describe product packaging, marketing, or digital rights and do not fit easily into the MARC records that populate library databases. Nevertheless, this information may be useful elsewhere in a library systems architecture to search, browse, evaluate, or manage access to products obtained from publishers.

4.1 Should RDA be used instead of AACR2?

The MARC records produced by the translation conform to the Anglo-American Cataloging Rules (JSC-AACR 1988) because of its widespread use in English-speaking library communities, but the discussion in the previous sections showed that it is sometimes difficult to give machine-processable expression to some of the newest concepts in this aging standard to some of the concepts introduced in ONIX 3.0. But if RDA, the intended replacement for AACR2, were the target of the translation instead, some of the conceptual distance between ONIX and MARC could be diminished.

For example, consider again the problem of describing a print antecedent to an e-book discussed in Section 3.3. This strategic relationship is described in MARC with a free-text keyword *Print edition:*, which is added to the generic linking field 776, Additional Physical Form Entry. But in RDA, the MARC linking fields have been superseded by the Relationship Designators (RDA 2011, Appendix J), a controlled vocabulary with dozens of entries organized

as a hierarchically structured ontology for describing relationships among works, expressions, manifestations, and items that goes well beyond the expressive range of the MARC linking fields as well as the list of values from List 51 of Codelist 12 that populate the ONIX <ProductRelationCode>. The relevant portion is shown in Figure 4.3. As the red text shows, the relationship between e-books and their print antecedents is defined as an *equivalent manifestation* reproduced in a different format, analogous to facsimiles and reprints. Other manifestation relationships specify vertical relationships and accompanying material.

Equivalent manifestation
- Also issued as
- Mirror site
- Reproduced as
  - Digital transfer, electronic reproduction, facsimile, reprinted as

Descriptive manifestation
- Description of
  - Analysis of, commentary on, critique of, evaluation of, review of

Whole-part manifestation
- Contained in
- Contains

Accompanied by
- Issued with
  - Filmend with, on disc with

Figure 4.3 Some RDA relationship designators

Unfortunately, there are no dedicated fields for expressing most of these relationships in MARC. Instead, they must fit into existing fields that accept free text. To be RDA-compliant, the MARC target of the ONIX <RelatedProduct> for an e-book antecedent could simply be recast as *767 $i electronic reproduction of: $z 9781594200823.* But software processes that operate on the RDA-encoded MARC records would have to recognize that the data in these fields sometimes have a privileged status as controlled vocabulary. Yet despite this limitation, the use of RDA relationship designators in a map from ONIX could be viewed as a slight improvement over the AACR2 version because the RDA relationship designators assert with authority that the versions are equivalent and that the e-book is derivative.

Another problem that might be solved by RDA is establishment of a reference for electronic devices that deliver intellectual content, which was discussed in Section 3.3. The RDA developers acknowledge the difficulties involved in the description of this material and have outlined a possible solution in the RDA-ONIX Framework (Dunsire 2007). In particular, the Framework establishes a link between ONIX ProductForm codes and a simplified set of values

on dimensions similar to the 007 descriptors, thereby combining the componential analysis strategy of a MARC description with the enumeration strategy of ONIX. To further simplify the relationship, RDA disassociates the *content* of a work, such as a speech or recorded music, from the physical storage device, or *carrier*. Since the ONIX <ProductForm> codes closely correspond to carriers, they can be mapped to RDA by specifying values on the relevant dimensions. Table 4.1 shows some sample entries. The columns designate semantic dimensions similar to MARC 007 values. The sample category labels resemble the ONIX Codelist 150 glosses. The unnamed BaseCarrierCategories represent placeholders that could be eventually be populated by registered vocabulary.

Note: The table below illustrates combinations of specified primary values used to construct BaseCarrierCategories that currently have a counterpart in AACR, MARC 21, and/or ONIX. The categories shown are illustrative only; they are not intended to exhaust the possibilities for producing valid BaseCarrierCategories using the specified primary values for the targeted attributes in the Framework. Similarly, the sample category labels are intended simply to illustrate the kinds of labels that might be used to identify a category for a particular community.

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 1:1:8	■								■													■	book
BaseCarrierCategory 1:2:8	■									■												■	flipchart
BaseCarrierCategory 1:6:8	■													■								■	sheet
BaseCarrierCategory 2:6:3		■												■			■						filmslip
BaseCarrierCategory 3:3:1			■								■				■								microfilm reel
BaseCarrierCategory 3:3:3			■								■						■						film reel
BaseCarrierCategory 3:3:5			■								■								■				audiotape reel
BaseCarrierCategory 3:3:6			■								■									■			videotape reel
BaseCarrierCategory 3:3:7			■								■										■		computer tape reel
BaseCarrierCategory 3:4:1			■									■			■								microfilm cartridge
BaseCarrierCategory 3:4:3			■									■					■						film cartridge
BaseCarrierCategory 3:4:5			■									■							■				audiotape cartridge
BaseCarrierCategory 3:4:6			■									■								■			videotape cartridge
BaseCarrierCategory 3:4:7			■									■									■		computer tape cartridge

Table 4.1 Vocabulary mapped to descriptors in the RDA/ONIX Framework (from Kiorgaard 2006)

The RDA-ONIX Framework solution has several advantages over the direct mapping of MARC elements that I have described in this article. First, the link is to 007-like values in a simplified concept space, which is less vulnerable to alternative interpretation and is, at any rate, maintained by a trusted international standards agency. Second, the corresponding ONIX values are given at least a sketchy definition, which is endorsed by the ONIX maintenance agency, EDiTEUR. The result is an alignment of the ONIX code with RDA's binary values on a set of dimensions that distinguish among the types of physical storage media. But the concepts still need to be named in a natural language. Informally, these names can be supplied by the ONIX gloss, but a more robust solution would be the creation of a controlled vocabulary, which would also be maintained by a standards agency. In technical terms, the

numeric values shown in Column 1 in Table 4.1 would be converted to descriptive terms and entered into the RDA vocabulary registry (Hillmann et al. 2010), which, like an online thesaurus, manages definitions, translations, and revision histories by achieving consensus in a community of practice.

Nevertheless, the RDA/ONIX framework is still only the outline of a solution. It is not yet a realistic alternative to the complex mappings we have defined for AACR2-encoded MARC record. The vocabulary now registered is not precise enough to distinguish among values for the various kinds of audio discs or the different circumstances for accessing electronic content. And neither the AACR2 encoding nor the currently registered RDA vocabulary can preserve the distinction recorded in ONIX Codelist 150 for e-resources that are downloaded vs. those that are accessible from a link. Progress requires one of two steps. First, we could register the relevant ONIX vocabulary in the RDA/ONIX framework and make RDA exactly as descriptive as the latest version of ONIX. But we would have to defend it against those who argue that the fate of the RDA vocabulary should not be dependent on the release schedule of ONIX and that controlled vocabulary designed for interoperability should be parsimonious. Another alternative is to adopt the existing minimally defined RDA vocabulary and work with MARC standards committees to define a new field containing the ONIX ProductForm code. Both of these solutions would stem the loss of information when ONIX is mapped to MARC, but they imply that the barriers to progress are cultural or political, not technical.

4.2 Conclusion: a mixed message about RDA

In sum, most of the elements that comprise a bibliographic description can be mapped successfully from ONIX to MARC. Subjects, language of the content, publication dates, titles, identifiers,call numbers, editions, publishers, adience levels, and names can be mapped between the two standards with relatively little difficulty or loss of information. These elements have two key characteristics that distinguish them from entities that are difficult to map:

- *They can be processed algorithmically.* Most of the data values of the easily mapped element consist either of alphanumeric codes, or of text that is fixed and/or stylized and can be transferred to the target format with minimal processing. By contrast, the potentially useful RDA ontology for relationships, which can be deployed in the description of e-books, is difficult to exploit because it can be distributed across many fields in a MARC record and must be discovered by parsing free text.

- *They are semantically independent.* In other words, everything about a subject is described in a self-contained *subject* entity; everything about an identifier is described in an *identifier* entity; everything about a contributor is described in a

> *contributor* entity; and so on. But a physical description is problematic because critical information is coded in many elements in a MARC record, which necessitates a complex mapping from ONIX and record-level validation of the MARC target to ensure that the description is internally consistent.

In other words, RDA offers the promise of only incremental improvement in the definition of entities—partly because the mapping to MARC is already reasonably successful without the need to invoke RDA and partly because the genuine RDA innovations are still incompletely realized and difficult to process in a MARC carrier. RDA's rich vocabulary of relationships present some opportunities for an improved mapping between ONIX and MARC, but we must acknowledge the fact that relationships that are important for the library and publishing communities are simply different. Publishers have the arguably simpler task of establishing an inventory of the products available in the marketplace, which are associated with useful methods of promoting, selling and delivering them, while libraries describe products of intellectual endeavor as they evolve over time. Nevertheless even in these different contexts, there is much common ground. For libraries as well as publishers, it is important to note that a book is written or illustrated by someone; that it has been published or made available by an institution; that it has a subject; and that it has been reviewed, summarized, or abridged. Wherever RDA makes it easier to express these relationships, it has made good on the promise to shorten the distance between the two most important standards for bibliographic description.

Acknowledgement

The design and implementation of the crosswalk described in this article represents a team effort by my OCLC colleagues Tim Savage, Renee Register, Shweta Rani, and Mel Beck. Renee now has her own consulting company and Shweta is an information technology professional in Bangalore, India. I am also grateful to Graham Bell, chief data architect at EDItEUR, whose careful reading of this manuscript caught many errors and produced many improvements in exposition.

References

All URLs were accessed on April 30, 2012.

Bell, Graham. 2012. ONIX for e-books. In "Find That E-book—Or Not: How Metadata Matters." *NISO Two-Part Webinar: Understanding Critical Elements of E-books: Standards for Formatting and Metadata*.(March 21). http://www.niso.org/news/events/2012/nisowebinars/ebooks_metadata/.

BIC (Book Industry Communication). 2011. http://www.bic.org.uk/.

BISG (Book Industry Study Group). 2009. "ONIX for Books 3.0: An introduction." BISG webcast (July 29). http://www.slideshare.net/bisg/bisg-webcast-onix-for-books-v30-introduction-1787290.

---. 2011. BISAC Subject Headings List, Major Subjects. http://www.bisg.org/what-we-do-0-136-bisac-subject-headings-list-major-subjects.php.

Coyle, Karen. 2012. "Taking Library Data from Here to There." NISO/DCMI joint webinar (February 22). http://www.niso.org/news/events/2012/dcmi/library_data/.

DCMI (Dublin Core Metadata Initiative). 2011. *DCMI Metadata Terms*. http://dublincore.org/documents/dcmi-terms/.

DOI (Dibital Object Identifier). 2011. The DOI® system. http://www.doi.org.

Dunsire, Gordon. 2007. "Distinguishing Content from Carrier." *DLIB Magazine*, 13,12 (January/February). http://www.dlib.org/dlib/january07/dunsire/01dunsire.html.

EDItEUR. 2009a. ONIX: About Release 3.0. http://www.editeur.org/12/About-Release-3.0/.

———. 2009b. *How to Describe Sets, Series, and Multiple-item Products in ONIX 3.0*. http://www.editeur.org/files/ONIX%203/ONIX_Books_Sets_and_Series_3.pdf.

———. 2009c. *How to Describe Digital Products in ONIX 3*. http://www.editeur.org/files/ONIX 3/ONIX_Books_Digital_Products_3.0.pdf.

———. 2009d. ONIX ISTC Registration Format. http://www.editeur.org/106/ONIX-ISTC-Registration-Format/.

———. 2010. ONIX for Books Code Lists Issue 12. (29 October). http://www.editeur.org/files/ONIX%20for%20books%20-%20code%20lists/ONIX_BookProduct_CodeLists_Issue_12.html.

Godby, Carol Jean. 2010a. "From Records to Streams: Merging Library and Publisher Metadata." In *DC-2010: Proceedings of The International Conference on Dublin Core and Metadata Applications*, ed. Diane I. Hillman and Michael Lauruhn: 138-149. Dublin Core Metadata Initiative. http://dcpapers.dublincore.org/ojs/pubs/article/view/1033.

———. 2010b. *Mapping ONIX to MARC*. (Report and crosswalk produced by OCLC Research.) Dublin, O.: OCLC Research. http://www.oclc.org/research/publications/library/2010/2010-14.pdf.

Godby, Carol Jean, Devon Smith, and Eric Childress. 2008. "Toward Element-level Interoperability in Bibliographic Metadata." *Code4Lib Journal*, 2, 2008-03-24. http://journal.code4lib.org/articles/54.

Hillmann, Diane, Karen Coyle, Jon Phipps, and Gordon Dunsire. 2010. "RDA Vocabularies: Process, Outcome, Use." *DLib Magazine*, 16,1/2 (January/February). http://www.dlib.org/dlib/january10/hillmann/01hillmann.html.

IFLA (International Federation of Library Associations and Institutions). 2011. Functional Requirements for Bibliographic Records. http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records.

ISNI (International Standard Name Identifier). 2009. http://www.isni.org/.

JSC-AACR (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 Development of RDA). 2009. *RDA: Resource Description and Access. Prospectus*. http://www.rda-jsc.org/rdaprospectus.html.

Kiorgaard, Deirdre. 2006. *RDA/ONIX Framework for Resource Categorization*. http://www.loc.gov/marc/marbi/2007/5chair10.pdf.

LC (Library of Congress). 2011. A Bibliographic Framework for the Digital Age. http://www.loc.gov/marc/transition/news/framework-103111.html

---. 1999. "MARC 21 Format for Bibliographic Data." *MARC Standards*. http://www.loc.gov/marc/bibliographic/.

Martin, Kristin E., and Kavita Mundle. 2010. "Cataloging E-Books and Vendor Records." *Library Resources & Technical Services.* 54,4: 227-237.

Mitchell, Joan. 2010. BISAC-DDC Mappings. Presentation to Dewey Breakfast/Update session at ALA Midwinter Meeting, Boston, [Massachusetts], January 16. http://www.oclc.org/dewey/news/conferences/bisac.ppt.Penguin Group. 2011.

Penguin Classics. http://www.penguinclassics.com/.

PNE-RTG (Provider-Neutral E-Monograph Record Task Group). 2009. *Report*. http://www.loc.gov/catdir/pcc/bibco/PN-Final-Report.pdf.

RDA (Resource Description and Access). 2011. RDA Toolkit. http://www.rdatoolkit.org/.

Rust, Godfrey, and Mark Bide. 2000. *The <indecs> Metadata Framework: Principles, Model and Data Dictionary*. http://www.doi.org/topics/indecs/indecs_framework_2000.pdf.

Smith-Yoshimura, Karen, Catherine Argus, Timothy J. Dickey, Chew Chiat Naun, Lisa Rowlinson de Ortiz, and Hugh Taylor. 2010. Implications of MARC Tag Usage on Library Metadata Practices. Dublin, O.: OCLC Research. http://www.oclc.org/research/publications/library/2010/2010-06.pdf.

Wikipedia. 2012. Crosswalk (metadata). http://en.wikipedia.org/wiki/Schema_crosswalk.