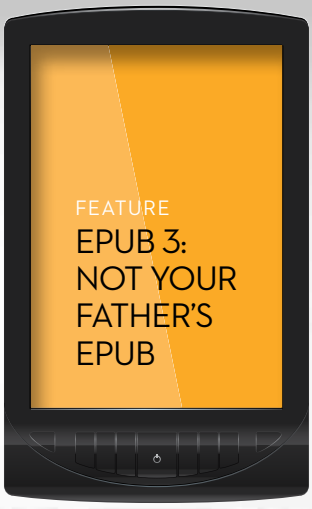
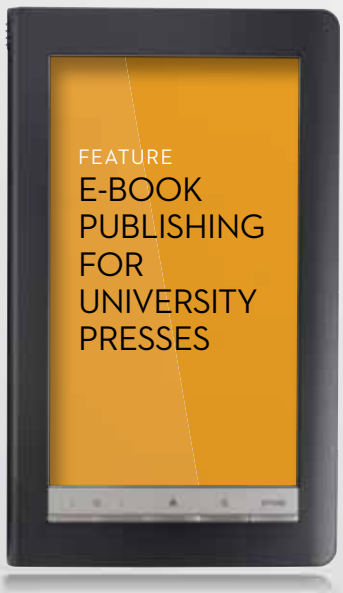
ARTICLE EXCERPTED FROM:

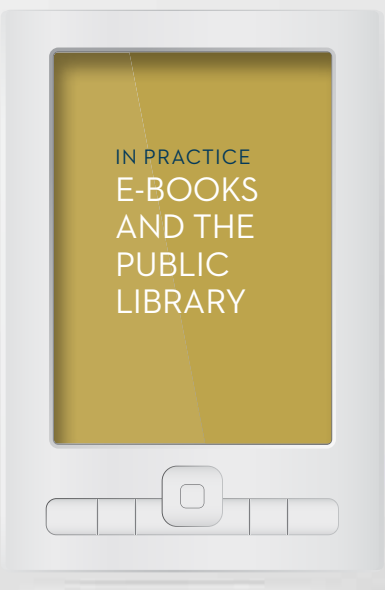


ISQ INFORMATION STANDARDS QUARTERLY

SPRING 2011 | VOL 23 | ISSUE 2 | ISSN 1041-0031

SPECIAL EDITION: VIEWS OF THE E-BOOK RENAISSANCE





NISO
How the information world CONNECTS

SP [SPOTLIGHT]

Mark Bide

STANDARD SPOTLIGHT:
ISBN AND ONIX FOR BOOKS

MARK BIDE

The Challenge for Standards in the E-book Supply Chain

The e-book supply chain is complicated—and is unlikely to get simpler any time soon. What do I mean by "the supply chain"? I mean the whole process that gets an e-book from author to reader—the only two ***really*** *important points in the whole chain. Without authors who willingly write and readers who willingly read, there would be no supply chain to worry about. But our primary focus in this article is those intermediaries who add value in the process from author to reader. Ask any author who has stood on a street corner trying to sell (or even to give away) copies of a manuscript to passers-by. The process which gets a book from author to readers adds value.*

In the physical world, the supply chain involves a number of different types of intermediaries coming together in different combinations: literary agents, publishers, typesetters and book manufacturers, distributors, wholesalers, retailers, and library suppliers. Now add the digital distribution chain alongside it, with its range of service providers and aggregators (sometimes the same companies, sometimes completely different ones). The process of shaping and re-shaping the digital supply chain is far from complete in 2011—but at least in the short term it isn't getting any simpler.

In both supply chains, there are myriad organizations that need to be able to talk to each other, to exchange information about the stuff that passes through the supply chain—in other words, to exchange metadata. And this is where standards come into the picture: they provide the common language that allows us to speak across organizational boundaries from machine to machine, oiling the wheels of this vastly complex enterprise, ensuring unambiguous communication and (to the extent possible) friction-free commerce.

All too often, discussions of metadata focus on a single application—discovery. Of course, in the digital world, discovery is in some ways a greater challenge than it used to be in the physical one because the only tool you have to bring readers to authors is metadata. Thus all online merchandising and marketing is about the metadata. So publishers are increasingly taking all aspects of their metadata seriously. While discovery is a particular driver, we cannot forget that high quality, accurate metadata lies at the core of all automated business processes. And efficient and highly automated processes lie at the heart of successful commerce on the network.

At the core of the metadata that drives those processes lies identity...

CONTINUED »

ISBN has quietly created the **backbone** of all of our standards and all our systems in the book trade.

The challenge of identity

The ISBN—which is arguably the most successful product identifier ever devised—was introduced to replace individual publisher's catalog numbers and to enable the first drive to electronic commerce. Without clarity of identity, it was not possible to use computers to manage the supply chain. The entire structure of EDI (electronic data interchange) standards on which an effective book supply chain has been built over the past 40 years has only been possible because of the implementation of the ISBN—distinguishing hardback from paperback, third edition from fourth edition.

ISBN has quietly created the backbone of all of our standards and all our systems in the book trade. This has been both our strength and our Achilles heel. The ISBN has enabled us to be sure that we are all "talking about the same thing," but its utility has been such that we have used it for all sorts of purposes for which it was never designed. The ISBN was (and is) intended to identify products in the supply chain. Look inside most publishers' systems, and you will find the ISBN used as a proxy to identify all sorts of things that are not products. It is not atypical, for example, for publishers to have a control on their cost ledgers such that it is not possible to incur costs on a publishing project without an ISBN. Whilst it may not need saying, a project and a product are simply not the same thing.

None of this mattered unduly when there was typically a close correlation between the "content" of the product and the "product" itself, and when the variety of products of any given "title" that could be made available was limited—maybe a hardback, a trade paperback and a regular paperback. The hardback ISBN was often used as the "master ISBN," to collocate (aggregate information about) this limited number of products (for example, in royalty ledgers). But note the sudden rash of quote marks in this paragraph. We are beginning to move into areas of uncertainty—areas where the meanings of words become uncertain and, critically, often mean different things to different people.

This is the type of ambiguity which is extremely threatening to the efficient operation of e-commerce systems. Computer systems are not good at resolving ambiguity and uncertainty. While we were dealing with physical products, the impact of this ambiguity was reasonably well controlled and rarely surfaced as a problem outside the walls of an individual publishing house (where it was rarely recognized for what it was). With the advent of the "e-book," however, the problem is suddenly becoming rather more acute.

The e-book and the ISBN

The answer to the question, "How do we identify our e-books?" seems very obvious. Use the ISBN. But it turns out it isn't quite as simple as that.

The first issue is the lack of clarity of what distinguishes one e-book from another e-book—at the product level. When the ISBN standard was last revised, e-books were still nascent. Although the current edition was published in 2005, the primary work on revising the text inevitably predates the formal publication (as anyone who has ever been involved in the creation or revision of an ISO standard will well understand). At that point, differences between e-book products were seen as analogous to the differences between a hardback and paperback—and the distinction that is drawn in the standard is between different technical file formats (the examples including PDF and HTML, as well as a number of file formats now obsolescent or obsolete). Perhaps understandably, what could not be foreseen at that time was that the development of the e-book market would not entirely mimic that of the physical book market, and that critical differences in the supply chain would make the application of a different ISBN

to each product much more challenging than it might have appeared in those early days.

There are several contributory factors to the rather unsatisfactory position in which we now find ourselves as an industry. Why unsatisfactory? Because what has previously been a reasonably consistently implemented standard, has evolved to a situation where there are widely differing practices in terms of ISBN allocation in different markets, between different publishers in the same market—and even sometimes between different parts of the same publishing house. Because of these differences in policy and practice, we are losing the certainty of identity that the ISBN normally affords us.

This is the first major upset that I can recall for the ISBN since the fierce debates with designers in the early '70s about the damage wrought to the artistic integrity of cover designs by putting barcodes onto books. This may be hard to imagine now, but it was very real then—at least until (in the UK at least) a dominant retailer announced "no barcode, no sale"; this closed the argument down very effectively. There are perhaps lessons to be learned from that experience.

Why is there a problem with e-book identification?

There are at least three major challenges with identifying e-book products with ISBNs.

The first is relatively straightforward and has already been mentioned. How do you distinguish one e-book product from another e-book product? The answer to this question has been provided in a set of Guidelines published by the International ISBN Agency. Although these guidelines may need to be further extended and nuanced over time, it appears that the general concepts that underlie them are proving robust.

However, although there is growing consensus at the theoretical level, there are still serious barriers to implementation. The first (and perhaps the most difficult) is that publishers do not know a priori exactly what products will be created from any given content, and cannot therefore easily pre-allocate ISBNs to the different products at an early point in the production lifecycle. ISBNs rather need to be available "on the fly" when the requirement for an additional product is identified. Unfortunately the creation and delivery of e-book products are typically not undertaken by the publisher, but by a digital service provider working on the publisher's behalf, or by an aggregator. There are no mechanisms available to the publisher—or to the service provider—to facilitate the issuing of these ISBNs at the appropriate point in the lifecycle (in other words, precisely when they are needed). One solution to this has been to allow these intermediaries to have their own prefix and to apply ISBNs themselves to publishers' products. But despite some successful implementations of this model (for example, by O'Reilly in their Safari online book product), it remains generally an unpopular option, particularly with publishers, not least because of the problems it creates for management of product metadata records. (It is not unusual for publisher systems to be unable to manage ISBNs issued by other publishers.)

Which takes us to the second problem that proliferation of products implies: this is frequently (pejoratively) referred to as "metadata bloat"—as if, somehow, metadata itself is growing out of control, a malign presence in the basement of the industry. Of course, the problem is that if you have a more complex world and more complex business, your metadata simply reflects that complexity. You don't simplify something by simplifying its description; with that approach you simply lose knowledge (data) about whatever it is you are describing, and this sort of data, once lost, is often impossible to regain.

This is not to suggest that there isn't a real problem here. Systems designed to manage a simpler world are often not appropriate for managing the sort of complexity that we are now facing. Many publishers' systems create metadata records for a new product by "cloning" the record of a related product and then editing the fields that identify the differences between the products. In the case of different e-book products, these may be very small differences. But now, instead of small numbers of metadata records for "the same" title, you have a growing number of individual records. And any time a change has to be made, there is no way of editing these records as a batch; each has to be individually edited, which is not only time consuming (and therefore expensive), but also error prone.

Although system solutions to this are in the development/deployment pipeline for the major vendors of publishing systems, it will be a while before they are anywhere near universally deployed. Quite apart from anything else, there is limited appetite for investment in systems at a time of considerable uncertainty. It is understandable that when the e-book market is doubling or tripling in size annually, grabbing market share and managing that growth takes precedence over any efficiency there may be in the better management of metadata. And, of course, it has to be recognized that the current explosion of different non-interoperable e-book products may be a passing phenomenon, with some sort of convergence point in the middle distance—in which case, why expend effort on a passing phase?

So, while some publishers have continued to recognize the importance of managing their different lines of product by effective identification, we have seen others deploying a single ISBN for all e-books (sometimes dubbed an "eISBN"). The one thing that is for certain is that there is no such thing as an eISBN—even if people are using a 13-digit number that looks like an ISBN)—because this identifier doesn't identify a product (the only class of entity that an ISBN can be used to identify) but

CONTINUED »

rather a class of products sharing the same content but different product attributes. Unfortunately, the idea of the eISBN has even reached the world of MARC cataloging (where its use has been promoted, despite the fact that it doesn't exist!).

And it isn't that these two positions (one ISBN for each e-book product, or one for all e-book products) are the only two models being followed. The reality is much more complex than this, with almost every imaginable practice being followed (including some publishers who persist in identifying their e-books with the hardback ISBN).

So, what we are faced with is an industry that has slowly seen its primary identifier system—once the flagship of identifier standards—slide into a chaos of incompatible practices and "workarounds." While I remain optimistic that we will establish international agreement on the implementation of ISBN, cleaning up the aftermath of the inconsistencies of the last four or five years is a different matter.

ISTC—an answer to the problem?

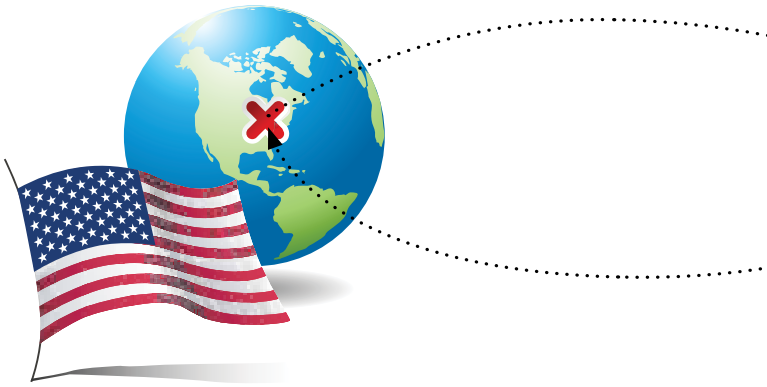
A contribution to resolving the challenge of collocation in the e-book market—drawing together the multiplicity of products containing the same content—could lie with broader application of the International Standard Text Code. This standard identifier, from the same ISO Committee which looks after the ISBN, is for identifying textual works—the abstract "content" rather than any specific manifestation of that content in a particular product.

However, this is another standard that is finding it hard to generate significant market traction. There are numerous possible explanations for this, but on the basis of recent research with publishers the requirement for a standard work identifier seems pressing. This apparent mismatch is deserving of further exploration.

E-books and ONIX

Having thought about some of the identification challenges we face, I will turn to the wider metadata picture. And for EDItEUR that means ONIX, and specifically ONIX for Books. The roots of its development lie in the 1990s, with the recognition by the Association of American Publishers (AAP) that there was a growing need for publishers to be able to communicate "rich product metadata" to online booksellers in an XML messaging format. The first release of ONIX was developed by the AAP, and the standard was then passed for long term governance to EDItEUR where we have managed it ever since. ONIX for Books 2.1 (released in 2004) has been widely deployed around the world; it is a credit to its designers that recent deployment in Japan has required minimal amendment to the standard.

However, it became clear about four years ago that ONIX required a major upgrade—and although in ONIX for Books 3.0



we have attended to a number of other weaknesses identified in earlier versions, the major driver behind the upgrade was the need to improve the capability to describe e-books. We undertook a major overhaul of the standard, with the approval and indeed encouragement of our International Steering Committee that represents the very large number of ONIX for Books implementers worldwide. However, release 3.0 does represent a significant challenge because in order to achieve the requirements identified by our users, we deliberately chose an upgrade structure that was not backwards compatible.

This creates a challenge for early adopters: why implement a messaging standard which no one can yet receive from me or send to me? Particularly when in ONIX 2.1 I can do 80% or more of everything I want to do...

This situation helps to explain the much slower progress towards release 3.0 implementation than hoped for when we launched it two years ago. We are now beginning to see more widespread implementation with the first major trade publishers following the pioneers in the academic market. I am optimistic that this will reduce the numbers of times that people ask me: "Why can't we describe this in ONIX?" when what they mean is "Why can't we describe this in ONIX for Books 2.1?" Because then I can stop giving the slightly frustrated answer, "Because you haven't implemented ONIX 3.0."

However, one challenge will keep recurring unless we create a significant change; ONIX for Books manages records at the product level. And here we return to ISBN country. We have recently once again been asked, "How do you describe more than one product in a single ONIX for Books record?"—a question to which the answer is (for the questioner) frustratingly clear: "You cannot." ONIX for Books is (and always has been) about product description—and a product (by definition) can only have one set of ONIX descriptors. A group of non-interoperable e-book products, with variations in file format, technical protection, usage limitations, hardware, or software requirements, and so on, remains a group of products and must be described with a group of product metadata records, notwithstanding that they manifest the same content.

About two years ago, recognizing and understanding the requirement for description of product groups, we put

There is considerable divergence in ONIX messaging practice, even within, for example, English language publishing. Messages from a US publisher cannot be interpreted by a recipient in the same way as messages from a UK publisher. This was fine while markets were organized nationally, but is posing an increasing challenge as the market and many of the key players in it become global.

a proposal for solving this dilemma to the ONIX for Books National Groups—the organizations represented on the ONIX for Books International Steering Committee (our governance group)—and it was unanimously rejected.

So we continue to face something of a quandary. Our pragmatic driver in standards development is to meet our stakeholder requirements; but our constraint is that our stakeholders find a degree of consensus. Part of our role is to facilitate that consensus—but that can be difficult when attitudes have become so polarized.

Growing pains

It is perhaps inevitable that the fundamental changes to the book industry that the "switch to digital" represents will be accompanied by some apparent lack of coherence when seen from the point of view of an organization whose role is to provide standards support. We are a long way from an understanding of how—or indeed whether—the shape of the market will settle down. However, one thing is becoming increasingly clear: markets are becoming increasingly global.

ONIX for Books has always been organized on the assumption that local implementations would vary from country to country and that "best practice" guides would be created at a national level. As a result, there is considerable divergence in ONIX messaging practice, even within, for example, English language publishing. Messages from a US publisher cannot be interpreted by a recipient in the same way as messages from a UK publisher. This was fine while markets were organized nationally, but is posing an increasing challenge as the market and many of the key players in it become global.

Our response has been to launch the first ever set of international best practice guidelines for ONIX for Books. While these will undoubtedly need to be supplemented locally—particularly in the physical book supply chain, where local practices will continue to need to be supported—we are optimistic that we can begin to resolve some divergences which have not been driven by any real differences in requirements, but simply by habit. The switch to ONIX for Books 3.0 is a real opportunity to improve consistency.

This is an essential step towards achieving another of our targets: more effective compliance. It is another commonly heard complaint that, "No one implements ONIX in the same way."

There are at least three possible explanations for this:

1. Inadequate or imprecise documentation, either from EDItEUR or from national groups
2. Imperfect implementation, based on developers not following documentation—either "guessing" at what things mean or because of a need to work around system inadequacies
3. Demands of powerful individual players in the market for customized data feeds, which are difficult to resist by smaller organizations (or even larger ones anxious to get their products to market) but which lead to a fracturing of the standard

To the extent that the first of these is in our own hands, we are doing what we can to improve documentation through the publication of the international best practice guidelines.

We can also help to some extent with the second of these challenges by giving direct support for implementations to our members (something we are offering on an increasing basis) and by publishing improved compliance testing tools—exemplified by our work on a Schematron schema for ONIX for Books 3.0, which enables users to validate messages against a much wider variety of parameters than either a DTD or simple XML schema.

The last challenge is more difficult. Ultimately, compliance is a peer-community challenge more than it is a central "enforcement" one. We cannot act as policemen; we can only exhort all those who implement our standards to be more forceful in driving out non-standard implementations—otherwise, the cost savings available through the implementation of standards can never be optimized.

CONTINUED »

Major recipients of ONIX data have a critical role to play here in encouraging data providers to adhere to broadly accepted standards—and to best practice—rather than demanding idiosyncratic proprietary interpretations. And we are keen to help these major recipients in their interpretation of the standards under our care, so our stakeholders can avoid costly recipient-specific metadata.

Meeting the challenge of convergence

The issue of the requirement for convergence arises throughout this article. In an increasingly global market, we need convergence between different organizations, and between different countries, in the way they implement the same standard (whether we are talking about ISBN or ONIX for Books). But convergence is going even further.

We are seeing convergence between requirements for ONIX for Books and those for what have traditionally been called ONIX for Serials messages. This family of messages—designed for communication within the library supply chain—was always exclusively focused on journal subscription products. However, we have recently undertaken a substantial overhaul of these messages to allow for them to cover any type of content that is provided on a subscription basis—including e-books and databases—and indeed non-textual resources. Although we have no present intention of enriching the ONIX for Serials messages with the sort of detailed product information that can be communicated in ONIX for Books, the message of the market is clear: the tidy distinctions between books and journals are rapidly being broken down. (There is also nascent interest in using ONIX to communicate about subscription products in the consumer market.)

And this brings us to the two final points that I want to make about convergence. The first is between ONIX for Books and MARC. These two standards have developed in very different ways—for good reason. There is a marked difference between requirements for book marketing and requirements for book cataloging and the different standards reflect these. Nevertheless, there is a dawning recognition of the potential for closer collaboration "across the divide." The work that OCLC has done in developing the ONIX to MARC (and back again) crosswalks is symbolic of this, as is the Library of Congress use of ONIX to improve the efficiency of its CIP program. EDItEUR is a partner in a European project called Linked Heritage which started in April 2011; our role in this project is to find ways to bridge the gap between commercial metadata and the Europeana digital library. All perhaps slightly tentative first steps, but all pointing in the same direction.

The final convergence challenge is perhaps the most significant but at the same time even more challenging to address than the differences between ONIX and MARC: convergence between the different media. Now that they can all be "consumed" on the same electronic device, it is proving increasingly difficult to draw the clear distinctions that we once so easily made between different media types. As the channels to market converge, it is entirely unrealistic to believe that we can continue to ignore the challenge that standards convergence will pose for us. We are only at the very beginning of this process and the journey in front of us remains obscure; but it is a journey on which we need to embark sooner rather than later.

A simpler life?

I cannot see any real likelihood that things are going to get radically simpler in the immediate future. Nevertheless, I remain optimistic that the challenges that we are facing—complexity, compliance, and convergence—are all actively on the agenda. EDItEUR is working in ever closer collaboration with its members and with other standards organizations all around the world and in all the different media to find ways to resolve our common challenges. The next few years will continue to be very active ones in the standards community.

| SP | doi: 10.3789/isqv23n2.2011.06

MARK BIDE <mark@editeur.org> was appointed Executive Director of EDItEUR in January 2009; he remains a Director of Rightscom, the specialist media consultancy where he has worked since 2001. He is a Visiting Professor of the University of the Arts London.

RELEVANT LINKS

EDItEUR
www.editeur.org

Guidelines for the assignment of ISBNs to e-books
isbn-international.org/faqs/view/17

International ISBN Agency
isbn-international.org/

International ISTC Agency
istc-international.org

Linked Heritage
www.cyi.ac.cy/node/1094

Mapping ONIX to MARC [OCLC]
www.oclc.org/research/publications/library/2010/2010-14.pdf (report)
www.oclc.org/research/publications/library/2010/2010-14a.xls (crosswalk)

ONIX and MARC21
www.editeur.org/96/ONIX-and-MARC21/

ONIX for Books
www.editeur.org/11/Books/

ONIX for Serials
www.editeur.org/17/Serials/

Provider-Neutral E-Monograph Record (This July 2009 report refers to something called an eISBN, while making it clear it is not product-specific.)
www.loc.gov/catdir/pcc/bibco/PN-Final-Report.pdf

Safari Books Online
my.safaribooksonline.com/

Using ONIX with Cataloging in Publication (CIP)
cip.loc.gov/onixpro.html