June 2002

ONIX -- A Transforming Standard

Michael Holdsworth

Cambridge University Press, myh@cambridge.org

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

Recommended Citation

Holdsworth, Michael (2002) "ONIX -- A Transforming Standard," *Against the Grain*: Vol. 14: Iss. 3, Article 6.

DOI: <https://doi.org/10.7771/2380-176X.3979>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

ONIX – A Transforming Standard

by **Michael Holdsworth** (Press Business Development Director, Cambridge University Press) <myh@cambridge.org>

The new superstores of the 1990s (**Borders** and **Barnes & Noble** in the USA; **Dillons** and **Waterstones** in the UK;) discovered a simple truth. The greater the range of titles offered, the greater the sales. Critical to this growth was the serendipity of “found” sales — when a bookbuyer comes into a shop for one book and leaves with three. Superstores replaced cramped uniformity with spacious diversity; now **amazon.com**, the “world’s biggest bookstore,” expands those choices to every available title. Previously hard-to-find backlist is taking an increasing share of total sales away from “top-twenty” bestsellers.

Internet bookselling created almost overnight a need for better descriptive product information. It was a need that the publishing industry simply had not recognised before and which it had no way of satisfying. A new standard for the exchange of bibliographic “metadata” — industry-wide agreement both on the data structures, and on the messaging protocol — was urgently required. Enter **ONIX** — the fruit of parallel transatlantic initiatives, which brought together the **AAP**, the **Book Industry Study Group** and the UK’s **Book Industry Communications (BIC)** under the umbrella of **EDiTEUR**, the international book industry standards body. The first issue of the **ONIX** standard appeared in late 1999.

Full information on **ONIX** can be found at www.editeur.org.

ONIX represents the most transforming new standard in our industry since the **ISBN**. The **ISBN** may be a deceptively simple and

undervalued identifier, but it has for nearly forty years formed a stable business base for discovering, organizing, sharing, trading and profiting from content. The *lingua franca* of any standard builds and grows partnership and community — and there is perhaps no global partnership as unified by a single standard as the community of publishers, booksellers, library vendors, wholesalers, librarians and scholars.

ONIX is a standard written in **XML (EXtensible Mark-up Language)**, a near-relation of **SGML** and **HTML**. Simply put, this means that it is plaintext based, consensual, universal, independent of operating system, application or platform. Which makes it affordable — and, to some extent at least — as readable, and as easily generated, by people as it is by machines. Simple marker tags within the data (expressed as angle brackets) identify when individual data fields start (<) and end (>), and are named according to **ONIX** rules. Think of using **Styles (italic, bold)** in **Microsoft Word**. Thus:

```
<ISBN>0521645956</ISBN>
<EAN13>9780521645959</EAN13>
<ProductForm>DB</ProductForm>
<ProductFormDescription>Paperback</ProductFormDescription>
<DistinctiveTitle> Capitalism Russian-Style </DistinctiveTitle>
```

This is no more or less complicated than clicking on, for example, an *italic* style in Word, then clicking on it again when you want to switch it off. All Word is doing is placing invisible tags at either end of the text.

What you ask a computer to do with this tagged text, or how you map it into a database, is up to you. While this means that **ONIX** is first and foremost a messaging standard, many publishers, particularly those building bibliographical databases for the first time, have taken **ONIX** as the basic design architecture and data dictionary. This has certainly been the case at **Cambridge**, where a suite of customised product and bibliographical databases have been reconfigured to match the **ONIX** data structure. The issue is usually one of **ONIX** either splitting or lumping data elements that may not have been separated in the original databases. When these older databases run templated desk-top publishing outputs — for example, for printed publicity and catalogues, or form the dynamic source-data for the publisher’s Web catalogue or e-commerce, the necessary adjustments can require considerable care in the implementation.

Since its launch in 1999, wider possibilities of **ONIX** as a single messaging structure for product information have been recognised. While constantly retaining a commitment to backward compatibility (to enable usability across revised versions of the standard), **ONIX** has been extended to encompass videocassette/DVD “biblio,” and eBooks. Minor tweaks to the structure have been incorporated to gain adoption of the new standard, beyond the English-speaking world, in France and in Germany, taking **ONIX** International from a multinational to a multi-lingual context. Discussions are underway

continued on page 24

The Amazon Effect from page 18

[Read “The Virtual Approval Shelf: A Look Towards the Future?”, pg. 28”].

• **Stephen Pugh** of **YBP’s International Division** writes as an approval profiler with many years of experience with selectors in US and International libraries. Selectors, he argues, barely have time to consider the paper and electronic announcements they now receive — will anyone actually embrace “slips on steroids?” [Read “The Decline of the Poultry Selector,” pg. 32].

Some Final Ruminations:

In this speculation about the virtual approval plan, we’ve deliberately focused on what may be possible, perhaps stretching credulity. In our temporary suspension of

disbelief, we’ve ignored some very real questions, which we must at least acknowledge in closing:

- The timeliness of data is critical in new title selection, and many of the data sources discussed still rely upon a published copy of the book to capture cover images, tables of contents, etc. The best hope for improved timeliness rests with publishers, and the degree to which a critical mass of them implement **ONIX**, with as much information as early in the publishing cycle as possible. That **ONIX** data may then be imported into vendor systems, or accessed by a link to the publisher site.

- It would be possible to test the timeliness of extended metadata for selection very easily. Identify the 800 or so titles profiled in a typical week by approval

vendors. Search each of them in a defined set of extended metadata sources to determine what percentage link to sufficient information for a selection decision that week. Any library school students looking for an interesting project?

- Vendor systems such as **Gobi** or **Collection Manager** play a significant role in approval selection in many libraries. For these services to thrive in a virtual approval environment, their systems may need to become **OpenURL**-compliant — certainly as target resources, perhaps as service providers.

Perhaps the biggest question of all, though, is an economic one: Is it sufficiently in anyone’s interest to assemble all the pieces? If we build it, will they buy? 🐼

with book industry communications bodies and publisher/bookseller associations as far afield as Korea and Argentina.

National libraries, and in particular the **Library of Congress** and the **British Library**, have also expressed interest in using **ONIX** as a base data-set for e-CIP, and to accompany the legal deposit of print or digital material. National ISBN agencies are exploring **ONIX** as a core data load to accompany **SBN** issue.

Nor has the world of serials been neglected. One of the most powerful features of **ONIX** is its granularity. The data-set can be optionally applied at any level of content, from book series, to book title, to individual chapter (and, if required, to sub-chapter level — to subheadings, paragraphs, or even sentences). For books, this offers users true chapter metadata for the first time, with all the benefit that this will offer for digital or print-on-demand or pay-per-view chapter (or other fragment) sales over the Web, for digital course-pack assembly, or for permissions clearances. The parallels here with the traditional journal SGML header file are self-evident, and plans are well in hand for an extension of **ONIX** to manage serials at the journal title, journal issue and individual article level. There are a number of papers describing the **ONIX** for Serials initiative available as downloads from www.editeur.org. For the first time, **ONIX** can offer the real possibility of uniform and interoperable books and journals information.

It is **ONIX**'s inherent power to bridge content types and information communities which most engages its supporters. Notwithstanding the fact that the standard was designed mainly by publishers and booksellers to facilitate Internet trading, **ONIX** is just as relevant within library applications.

After all, why should scholars, researchers and students, now daily accustomed to the rich presentation of title information on **amazon.com** (and the other Internet retail majors) or on the better publisher sites, settle for any less from their library **OPAC**? Or to turn the question around, what data elements — a detailed book description, author authority and biography, full structured table-of-contents, jacket illustration, other-books-by-this-author, full LOC data and CIP, a selection of the standard BASIC/BIC/LOC subject-codings and **Dewey**s, sample text, book reviews, prizes and awards won, sound-clips, author-interviews, market rights and territorial restrictions — which of these could possibly be deemed *inappropriate* for library patrons? And why stop there, why should not more technical information — book dimensions, weights, font and production data — be made available for logistics people or design aficionados? We know that

AUTHOR BIO

Michael Holdsworth is Press Business Development Director at Cambridge University Press. He has wide-ranging global responsibilities for change and new technology issues at the Press, spending much of his time on initiatives for systems development, content and rights management, production workflows, Web-based development and the supply chain.

library patrons, faced with a minimalist **OPAC** title entry, will often go in any case to publisher and retailer Websites for just this sort of extended information. Why should it not be our mutual intention as an industry to seek to move all descriptive data into one common similar look-and-feel throughout bookstore, publisher, wholesaler, library vendor, library Websites and catalogues — an interface which is not just the skeletal *identifier* that is a MARC record, but a genuine *tool* for discovery and research? For books and for serials alike; for book chapters and for journal articles; for video, sound and e-content. **ONIX** provides that key.

Which is not, of course to say that an **ONIX** record need necessarily be as rich or packed as the previous paragraph might suggest. The majority of the data elements are optional, and it is expected that a core set of mandatory elements can be agreed, which can form the basis of industry accreditation (as “**ONIX**-compliant”). In the UK, **BIC** has had considerable success in promoting the certification of publishers as **BIC-Basic** compliant, able to deliver a core set of bibliographical and availability fields at a number of key points in the publishing process.

The big cultural change implied by the new **ONIX** is that it pushes responsibility for bibliographic data back upstream to publishers, and challenges everyone else downstream in the information supply-chain (booksellers, vendors, librarians) to trust those publishers to manage this properly. It remains to be seen whether publishers will rise to that challenge, and where this will leave those in the business of aggregation.

There can be no doubt that the current metadata supply chain is seriously flawed. Most databases are populated by unsystematic multiple feeds both from manual entry and from data “cascades” as computers (at the bookstores, at the aggregators, at the vendors, in the libraries) poll each other to fill gaps in their data-sets by combining data, often applying arcane rules about which entry should trump or overrule another. Where errors do arise it can be extremely difficult to trace the cause.

Less error-prone, but hugely expensive and wasteful is the process of physical “book-in-hand” processing. This can extend to any number of processes: manual transcription of title-page and front-matter, retyping of table-of-contents, physical weighing and measuring, digital scanning of jacket art or back-matter, author-authority research, page and

illustration counting/analysis and so on. We sometimes wonder how many times a **Cambridge** title is subjected to these identical processes within the 150 or so countries in which we sell our books. And self-evidently the book-in-hand process, being manual, is susceptible to error; and by definition, since it requires the actual book, it is always late, often too late to be much use for the selection process.

It is equally self-evident that all this information is available from the publisher — the horse's mouth — unmediated, and without introduced error, and at the earliest possible point in the production workflow. **ONIX** messaging, either directly from the publisher to the user, or channelled unedited through data aggregators, opens up a world of new opportunities for a faster, cheaper, more accurate information supply-chain.

The step from a rich **ONIX** data-flow to “virtual approval” is straightforward. **ONIX** provides an effective surrogate for front-matter, and allows for substantial sample text elements. Most academic publishers have long ceased being coy about using sample chapters (or even, in a controlled manner, the full text) on the Web for marketing within their own catalogue sites. **Cambridge** allows library vendors direct access to PDF chapters on its own site, fully integrated with the vendor's own collection management offering, freely accessible both to the vendor's own approval selectors, and to their librarian customers. And all of this can be made available months before the first copy of the book is even printed.

In 2000, **Cambridge** launched its path-breaking **DataShop** facility: <http://datashop.cambridge.org> which provides a easy-to-use, self-service data delivery service with free access to anyone (booksellers, researchers, librarians, data-aggregators, individuals) who wants to benefit from rich, 100% current, authoritative information on our publishing. **DataShop** offers a range of easy downloads from simple price-and-availability, to jacket files, to full-blown **ONIX**; anyone reading this as an **ONIX** novice is welcome to try the service and email themselves a trial file or two! **Cambridge** sees clear advantages in moving away from pushing data out, to having users pull information in, customised to their personal needs, and scheduled according to their own individual interval preferences.

continued on page 26

The Virtual Selection Process – The ILS Perspective

by Ted Fons <tfons@iii.com>

Introduction

Book reviews, book jacket images, synopses, tables of contents, author biographies, publisher profiles, up-to-the-minute vendor inventory snapshots, first chapters, and many other tools of the selection process can be made available to library selectors and collection development staff via the Web in an integrated way. Integrated in the sense that while the selector is considering a new title, these extended metadata elements can be consulted from a single integrated library system interface and a Web browser.

Imagine a selector retrieving a title under consideration for purchase inside the **Integrated Library System (ILS)**. At the time that the title is retrieved, the **ILS** can build a set of hyperlinks to all of the resources mentioned above and launch the Web browser when any of the links are selected. To determine the scope of the title under consideration, the selector might want to click on the link for a synopsis offered by Informata. The selector might be curious about the background of the author, so a **Bowker's Books in Print** link is selected to retrieve an author biography. To dig deeper into the intellec-

tual content of the work, the selector might refer to the table of contents, or the first chapter of the book offered by **Syndetics Solutions**. For a title published outside of the United States, **Bowker's Global Books in Print** on the Web might be used. When other sources fail, a **Google** search might be used to cast a wider net. Because this selector is involved in a cooperative collection development project, the next choice might be a direct search link to the Web-based catalog of a nearby college, which is also the selector's collection development partner institution. Finally, when the selector is getting closer to a selection decision, **Baker & Taylor's** Web-based inventory service might be consulted to determine if the title is currently available from **B&T's** local distribution warehouse. All of these sources are available with content-specific hyperlinks from within the **ILS**. Many **ILS** vendors are offering links from the patron-focused Web **OPAC**. **Innovative Interfaces** is offering these links in both the patron and staff interfaces.

Sound judgment, subject expertise and experience are still among the most important characteristics of a selector, but con-

venient access to the enormous variety of resources now available on the Web adds efficiency to the process of making an informed selection decision. Linking to these sources from the **ILS** makes sense because the selector already uses the **ILS** to check for local holdings and may refer to the acquisitions system to check on the availability of money for new purchases. The scenarios described above work for a both an approvals review workflow where bibliographic records have already been loaded onto the **ILS** for bibliographic control and payment purposes, and in the firm order workflow where the selector is using the **ILS** to determine local holdings. In the case where the title is already held by the library, a trained selector can refer to circulation statistics to make a decision about ordering additional copies of a title. In the case where the title is not held locally, the **ILS** can also make connections to coop-

continued on page 28

ONIX

from page 24

There are two larger questions, both addressed I am sure within this special issue of **ATG**. First is whether booksellers, vendors and libraries will ever want to take individual publisher feeds, rather than rely on aggregators. The **Cambridge** view is to hope for a mixed economy, where a few (maybe a dozen in each country) of major publishers will deal with a few major users (again, maybe a dozen), while most data exchanges will continue within aggregated services. Second is whether the target, in the library sector, for such new publisher **ONIX** services, is actually the vendor's approval specialist or the individual acquisitions librarian.

There is one final dimension — which I will call “intrinsic” metadata. Most (nearly all...) book MARC or **ONIX** elements are at the same time metadata *and* content. Title, subtitle, series-name, author-name would be obvious examples, but the same is true of table-of-contents, book description (as jacket “content”) and all the text on the copyright/imprints page (dates, imprint, CIP and all its LOC/BL coding and

components). These are the *content* elements which are transcribed into *metadata* at book-in-hand, but they are actually one and the same thing.

Cambridge, in common with many publishers, is now “typesetting” using XML, so that each book becomes, effectively, a consistently structured database, with tagged elements that recall the structure of **Microsoft Word** mentioned earlier in this piece. Development work is in progress at **Cambridge** electronically to extract the elements needed by **ONIX** as a routine part of the production process, thereby obviating any duplication of data entry at the publisher end and eliminating transcription error. With **ONIX** being “harvested” dynamically from work-in-progress, we can now envisage a situation in which a proofing correction to digital “galleys” made by a copy-editor — perhaps to a chapter title — would be reflected in real time on **Cambridge DataShop** outputs, and via **ONIX** messaging instantaneously to the wider community of our bibliography consumers worldwide. Richer data, months earlier than is possible today, and with 100% accuracy (or at least a 100% match with the book itself!).

Savvy publishers will soon be able to offer this scale of virtual experience, either on their own Website, or via **ONIX** feeds supplemented with sample text access, with these feeds enriched with all the LOC and BL CIP components and author-authority. What value-added elements will then be left for the vendors and aggregators? Even subjective elements, such as subject-coding (on any choice of industry-standard set), may be better left to the publisher's specialist academic editors. From both the publisher's and the data consumer's perspective there seems nothing to lose here but redundancy, delay and potential error. **ONIX** has the potential to free up resources — within libraries for better acquisitions management, and within vendors for better service — a win-win all round.

At the same time, in the online retail environment for which it was always intended, **ONIX** allows Internet bookstores better to recombine and present families of products — by author, by subject, by theme, by level — in their continuing pursuit of that perfect marketing match: the right book for the right customer. 🐾