Zen and the Art of Metadata Maintenance

John W. Warren
ORCID iD: http://orcid.org/0000-0002-4882-7205
Researcher iD: M-5309-2014

This is a copyedited version of an article published by Journal of Electronic Publishing in June 2015. The version of record is available online at DOI: http://dx.doi.org/10.3998/3336451.0018.305

Abstract:

Metadata is the lifeblood of publishing in the digital age and the key to discovery. Metadata is a continuum of standards and a process of information flow; creating and disseminating metadata involves both art and science. This article examines publishing-industry best practices for metadata construction and management, process improvement steps, practical applications for publishers and authors such as keywords, metadata challenges concerning e-books, and the frontiers of the expanding metadata universe. Metadata permeates and enables all aspects of publishing, from information creation and production to marketing and dissemination. It is essential for publishers, authors, and all others involved in the publishing industry to understand the metadata ecosystem in order to maximize the resources that contribute to a title’s presence, popularity, and sale-ability. Metadata and the associated processes to
use it are evolving, becoming more interconnected and social, enabling linkages between a broad network of objects and resources.

**Keywords:** metadata management, publishers, ONIX, content, information, Linked Data, process improvement, social metadata
The Magnitudes of Metadata

In the short film *Powers of Ten* by Charles and Ray Eames (1968), we begin by viewing a couple picnicking along the lakeside in Chicago while enjoying a beautiful day. Then we soar by an order of magnitude, a rate of one power of ten per ten seconds outward, above the lake, above the city, reaching the outer edges of the universe, 100 million light years away. Zooming back to earth, we move into the hand of the sleeping picnicker—and are transported with ten times more magnification every ten seconds. We end up inside a proton of a carbon atom within a DNA molecule in a white blood cell.

Metadata is both a universe and DNA. Often defined as “data about data,” metadata can be defined more broadly as information that describes content, or as an item that describes another item. In this sense, metadata is more than just basic descriptive data: a book review about a scholarly monograph becomes part of the monograph’s metadata, while the book review itself is content with its own metadata. The book becomes, in the broadest sense, part of the metadata of each book or article it references; the author forms part of the book’s metadata, and the book becomes part of the author’s own metadata.

Metadata is the lifeblood of publishing in the digital age and the key to discovery. Metadata permeates and enables all aspects of publishing, from information creation and production to marketing, discovery, dissemination, involvement, and impact. This article argues that it is essential to understand the metadata ecosystem in order to optimize a title’s presence, popularity, and sale-ability, while reminding users that metadata is evolving—shifting from a publisher- or institution- assigned process to being dispersed into social and retail spaces where diverse actors and stakeholders have different views of and
intentions for the content. Even as metadata becomes an increasingly strategic function of the publishing enterprise, metadata is becoming interconnected, crowd-sourced, and employed in new ways across the information spectrum.

“Seeing Standards: A Visualization of the Metadata Universe,” created by Jenn Riley and Devin Becker (2009–10) of the Indiana University Libraries, presents a striking array of the communities, domains, functions, purposes, and standards of metadata within the cultural heritage sector. Museums, libraries, the information industry, and archives are communities that both create and use metadata. Musical materials, cultural objects, datasets, and scholarly texts, among others, are domains for which these communities use metadata standards to describe their materials.

(Reproduced from Riley and Becker 2009–2010, under CC by NC SA 3.0)

Zooming within the domain of scholarly texts, we note a plethora of standards, such as ONIX, used within the book trade. This is a sampling, as Riley and Becker (2009–10) point out, “the standards represented here are among those most heavily used or publicized in the cultural heritage community, though certainly not all standards that might be relevant are included,” but the visualization is remarkable in revealing their inter-relationships.
The goal of book metadata, whether the text is scholarly, trade, or self-published, is to make a book discoverable on the myriad of websites and databases used by libraries, consumers, researchers, and students, and impel the book’s purchase or acquisition. At its most basic level, metadata is information: product basics such as the title, author, and jacket cover; product description such as catalog or back-cover copy; product details including identifiers (i.e. ISBN), price, page counts, and carton quantity; additional information such as author bios, blurbs, reviews, awards received, rights information, keywords, table of contents, and price in other currencies; and even more information, from book trailers, sample chapters, and excerpts to related books.

Metadata is information, and it is also information flow. Publishers create metadata and send it to retail and other partners, either directly by the publisher or through intermediaries, including metadata management companies and distributors. Retailers parse the metadata in order to display the product to their customers, whether online or in store.
In the ideal system, metadata is never altered between leaving the publisher and reaching the retailers, represented by the clean lines above, while information is shared both ways between partners. But metadata is messy; the figure below shows how it actually works. What were once two-way arrows are now one-way arrows, and the red arrows show where metadata is being modified downstream (BISG 2013).
In metadata’s flow through the information spectrum, many hands and systems touch it, and it is hard to tell where problems originate; it is created or altered by other parties in the publishing ecosystem for their own needs. Yankee Book Peddler (YBP), for example, has long relied on their own subject experts to catalog and codify metadata from final, printed books to meet the demands of their academic library customers.

The Rise of Metadata

In the fourteenth and fifteenth centuries, before Gutenberg’s invention of movable type that led to the birth of printed books, producers and collectors of manuscripts kept catalogs of their treasures as books became a part of international trade. In the early age of print, libraries, collectors, and book dealers managed their inventories in correspondence and account books; the Frankfurt Book Fair emerged by 1475 as an important trading place for books.
the early sixteenth century, the development of the book trade led to the rise of printer’s guilds and to the creation of standards, such as the book’s title page; all of these factors led to the creation of metadata although it wasn’t yet referred to by that name (Pettegree 2010, 18, 39, 66, 79). Kwakkel (2014) compares the use of shelfmark systems within medieval manuscripts, catalogs, and even “handheld devices” used to locate volumes, often chained and placed on lecterns in the medieval monastic library, to a modern GPS device. The term “metadata” was coined by Bagley (1968) in advancing concepts in computer programming languages. Metadata has been standardized, codified, and indexed primarily by information professionals for at least the last hundred years; while this process has accelerated in the information age, metadata is no longer the sole purveyance of publishers and librarians alone (Gilliland 2008, 1). User-created metadata has exploded through social-bookmarks, reviews on Amazon and Goodreads, and in tags on LibraryThing, among others.

(Source: Google Ngram viewer, https://books.google.com/ngrams)
Google’s Ngram viewer allows researchers to search keywords or short phrases (ngrams) and visualize data on the usage of these terms and phrases over time, from the corpus of titles in Google’s book digitization project. The image above shows the rise in the term metadata mentioned within books in the English language from 1960 to 2008. The Ngram database itself presented considerable metadata challenges. Google digitized 15 million books out of 130 million titles published since the late 15th century; five million of these digitized books were selected for the Ngram database, a subset comprising roughly 4% of all published books from the period and totaling 361 billion words in the English language alone (Michel et al. 2011a, 176). Compiling metadata from thousands of sources, including library catalogs, retailer and wholesaler databases, and publisher submissions, Google discovered that many of the sources, not surprisingly, were riddled with errors or inconsistencies, requiring extensive clean up efforts (Aiden and Michel 2013, 83; Michel et al. 2011b, 3–17). Meanwhile, an opposite discovery problem was required from the Ngram—a prerequisite to not display or provide title-specific bibliographic data due to uncertain legal issues around copyright (Aiden and Michel 2013, 85).

When Amazon set out to catalog and offer for sale the world’s largest selection of books in 1995, the nascent online retailer underestimated the bibliographic issues; describing books in an online commercial environment was more difficult than anticipated. At the time, publishers and distributors utilized a supply-chain approach to metadata, at least in electronic form (excluding printed catalogs and brochures): fixed, limited fields, truncated titles, ISBNs that were often unreliable, and a lack of cover images was the norm. Quickly discovering that cover images were a key component to maximizing sales, and realizing that most publishers were unable to supply them, Amazon devoted resources to swiftly scan covers, with the implicit or explicit blessing of publishers, for online
display. Ingram and Baker & Taylor also began to expand their metadata efforts, but by 1999, it was clear that there were significant shortcomings. Incomplete and inaccurate metadata as well as conflicting requirements took their toll on emerging e-commerce (C. Aden 2014).

Carol Risher, then-Vice President for Copyright and New Technology at the Association of American Publishers (AAP), brought key players together in an unprecedented meeting, in July 1999; it included more than sixty publishers, online booksellers, wholesalers, and other key stakeholders. She may have been motivated, at least in part, by complaints about publisher metadata from her stepson David Risher, who at the time was Senior Vice President and General Manager at Amazon. As a result of that meeting, AAP led a fast-track project that resulted in the development of the first version of ONIX in January 2000, a new standard that enabled publishers to supply “rich” product information to Internet booksellers (Hilts 2000). Cindy Aden, then-surnamed Cunningham, represented Amazon on the project committee. Aden notes that this coming together of publishers and retailers was momentous, as archrival retailers like Barnes & Noble and Amazon agreed it was better to compete on factors like price or availability instead of vying over basic metadata (C. Aden 2014).

ONIX—ONline Information Exchange—is a standard for representing book, serial, and video product information in electronic form. ONIX feeds are sent to book industry trading partners, including wholesalers (Ingram, Baker & Taylor); online retailers (Amazon, Powells); retail chains (Barnes & Noble); and databases (Library of Congress, Bowker Books in Print, Serials Solutions, OCLC/WorldCat). Enriched ONIX data can include hundreds of fields of data such as description, rights information, and reviews. EDItEUR, jointly with Book Industry Communication (BIC, based in UK) and Book Industry Study Group
(BISG, based in US), maintain the ONIX standard in cooperation with industry stakeholders.

ONIX is encoded as XML (EXtensible Markup Language); it follows a logical structure, and although it appears pretty technical at first glance, it is both human- and machine-readable. Here is a snippet of a typical ONIX record:

```
<Product>
    <RecordReference>1234567890</RecordReference>
    <Notification Type>03</NotificationType>
    <Product Form>BB</ProductForm>
    <DistinctiveTitle>Hezbollah</DistinctiveTitle>
    <Contributor>
        <ContributorRole>A01</ContributorRole>
        <PersonNameInverted>Levitt, Matthew</PersonNameInverted>
        <BiographicalNote>Matthew Levitt is a senior fellow and director of the Washington Institute for Near East Policy’s Stein Program on Counterterrorism and Intelligence</BiographicalNote>
    </Contributor>
    <LanguageOfText>eng</LanguageOfText>
    <NumberOfPages>368</NumberOfPages>
    <BISACMainSubject>POL012000</BISACMainSubject>
```

ONIX has evolved and is continuing to evolve with a goal towards improving the accuracy and richness of book metadata: after the ONIX 1.0 release in January 2000, version 2.1 was introduced in 2004, and ONIX 3.0 in 2009. The current standard, ONIX 3.01, was released January 2012. Many industry partners still do not accept ONIX 3, although EDItEUR/BISG plan to sunset ONIX 2.1. According to the BISG, ONIX 3 is more robust, nimble, and better suited for managing digital content (BISG 2013).

Publishers should strive to improve the quality of their metadata as well as increase the quantity of accounts to which they send metadata feeds directly in industry standard ONIX. Even as ONIX is industry-preferred, publishers must
nimbly deliver metadata using a variety of methods due to the diversity of trading partners, which vary in their market focus, business strength, and technical demands. Other common transfer arrangements include Microsoft Excel or comma-delimited file formats. Vendors also request cover files in a wide variety of formats and sizes. Each publisher will ultimately need to determine the optimal, or manageable, number of metadata elements, file formats, and image sizes that they provide to their trading partners. Best practices guides and code specifications are available on BISG and EDItEUR websites.

BISG’s guidelines identify 32 core metadata elements—of which 27 are listed as mandatory, albeit some mandatory only if physical products, some if digital, some in special circumstances (BISG 2014c). This does not mean, however, that trading partners will reject publisher metadata with missing “mandatory” elements. The majority of these core elements have various sub-elements and iterations. The full ONIX specification (release 26; July 2014) comprises 156 code-lists that include a whopping 4,105 different values, and that does not take into account the nearly innumerable iterations within those values, such as the 3,000-odd BISAC subject codes to be used for code list 26, value 10 (EDItEUR 2014).

Using the ONIX code to indicate audience is one example. For each title, the publisher selects one of the following audience codes from the ONIX Code List 28 (EDItEUR 2014; BISG 2014a), though, naturally, many titles will potentially appeal to more than one of these audiences:

- 01- General/Trade, for a non-specialist adult audience
- 02- Children/Juvenile
- 03- Young Adult
- 04- Primary & Secondary/Elementary & High School
- 05- College/Higher Education
- 06- Professional and Scholarly
- 07- English as a Second Language
According to BISG, the element most frequently in error within publishers’ metadata is the subject code (BISG 2013). This might seem odd—after all, one would expect publishers to know the subjects of the books they publish. Except when considering that the subject codes used, BISAC Subject Headings (in the US) include approximately 3,000 “minor” subject headings grouped under 51 “major” subjects; BIC codes (used in the UK), consist of 2,600 different subject categories arranged in 18 sections defining broad subject areas, plus approximately 1000 qualifiers used to refine the meaning of the subject categories (BISG 2014c, 100). That, and it may be an intern or a lower-level marketing coordinator that is assigning subject codes for a publisher’s titles. Publishers, for their part, also claim that BISAC subject headings are the most frequent element modified by metadata recipients (BISG 2012).

Today’s content creators and providers, including publishers and other information organizations, require staff trained and specializing in metadata management. Few publishers, however, are prepared for the necessary commitment of managing metadata, with continually evolving standards and an often inconsistent set of demands from trading partners. Thus, intermediaries have filled the gap, particularly for smaller and medium-sized publishers but often even for the Big Six.

“When Pete Alcorn and I founded NetRead in 1999, we emphasized the value of metadata as a marketing asset,” says Greg Aden (2014):

“We believed that our metadata engine, JacketCaster, would drive the creation of print catalogs, tip and buy sheets, press releases, and other marketing materials online, and this value-add would usurp the value of our ONIX generation capability. But due to the fundamental and essential need for getting the product to market, our resources were immediately pressed into service toward data integrity, increased focus upon ONIX
and its upgrades, and reliable, but flexible, data transmission and protocols. Diligent attention to the standard, continued support of other metadata formats, practices, and international demands turned a niche into a business supporting Big Six publishers, university presses, distributors, and small presses. With the arrival of digital assets, we offer additional services such as ebook distribution and conversion, but the heart of our application remains with metadata, and is architected around the ONIX standard.”

Metadata management serves more than a marketing function. When subject-level, audience, and age-range metadata are combined with sales data, it allows publishers to assess sales trends and emerging opportunities; profit and loss analyses that employ metadata on subjects, acquisitions editors, and other factors provide a granular view into performance that can improve decisionmaking. Metadata’s central role in the information organization includes administrative issues (rights and royalty information); interoperability (transmission of data between or within internal and external systems and standards); production (flow of products through creation to dissemination); marketing and distribution (data exchange with distribution, retail, and other vendors and partners); repurposing (chunking content into chapters and micro segments); learning (integration of content into learning management systems and databases, from Blackboard to SCORM); preservation (digital assets management, version control); and use (tracking, search logs, and e-book usage).

In the future, publishing programs, like the masters in publishing at New York University or George Washington University, would be wise to provide courses and even specialized tracks in metadata management.

**Metadata Matters**

It may seem intuitive that better metadata has a positive impact on sales; however, the precise amount of effort to expend on metadata improvement for a
given return on investment in sales and discoverability is an equation that will vary from publisher to publisher. A Nielsen Book (2012, 3) study on the impact of metadata on sales reported that books with complete “basic” data sold twice as many copies on average than those with incomplete metadata. The basic data included 11 required data elements to meet the BIC Basic standard. The study tracked UK sales for Nielsen BookScan’s top-selling 100,000 titles of 2011, but the results are clearly analogous for the US market as well. Titles with complete basic metadata and cover image sold nearly six times as many units as those with incomplete metadata and no image, and 56% more than those with a cover image but incomplete data. Interestingly, complete metadata had an even greater effect on boosting average sales in brick and mortar booksellers than for online retailers (Nielsen Book 2012, 4).

![Metadata: average sales per ISBN](image)

(Reproduced with permission from Nielsen Book, “The Link Between Metadata and Sales,” 2012, 3)

Nielsen Book’s study also analyzed the effect of enhanced metadata fields on sales; books with four enhanced elements—short and long descriptions,
review, and author biography—present on each record sold on average 1,000 more copies per title than those without these enhanced elements. In this case, the effect was strongest for online sales and was strongest for fiction titles, followed by trade non-fiction, specialist non-fiction, and children’s books (Nielsen Book 2012, 5–6). While it could be argued that larger, more successful publishers are likely to have better metadata and also more liable to attract better-selling authors and titles, it is hard to dispute that sales will improve with more robust, timely, and accurate information.

Mindful content creators would do well to develop and instill standards and principles for metadata maintenance. Metadata creation and management is a core activity and a shared responsibility, perhaps as important as the content itself. The adage “junk in, junk out” is especially apt concerning older backlist titles but can befall front list as well. Many publishers mitigate errors by checking and updating metadata from the final physical product, and some routinely check trading partner sites for metadata accuracy (BISG 2012; Baca 2008). Systems as well as processes should ideally be simple yet flexible and robust, which is a tricky combination. Manual fixes to metadata exports, such as manual alterations to the rights fields on certain titles, should be avoided at all costs, as the list of exceptions will inexorably grow.

Metadata management involves processes and procedures—events, roles, handoffs, and the like—and as such, can benefit from continuous improvement efforts such as those implemented using Six Sigma techniques. Publishers must understand as well as implement the basics, but they should never stop trying to improve both the quality and quantity of metadata. The effort never ends, whether exporting data directly to trading partners or via third parties such as Firebrand or NetRead; one of the paradoxes about metadata is that there is always something to add, improve, or clean up. Data standards change and new
fields become adapted and applicable. With a goal toward increased metadata process efficiency, publishers can start with “low hanging fruit,” relatively easy process improvements such as adding a brief description in addition to a longer description or modifying who enters a particular data element. Then they can move toward bigger projects, for example, like turning four different databases archiving metadata elements and sales information into one repository.

Mapping the metadata process is a step toward implementing improvements in accuracy and timeliness. Within the publishing organization, metadata is entered into a database or databases by various departments or personnel, and flows from the publisher via ONIX or spreadsheets through numerous systems to trading partners, ultimately to thousands of destinations. A metadata process map diagrams the flow of information, recording details such as what personnel or department inputs data, handoffs, data repositories, and to whom metadata is exported. A process map may reveal problem areas such as disconnects, bottlenecks, redundancies, rework loops, and decision delays (Pande, Neuman, and Cavanaugh 2014, 266–267). In the detailed view below, which represents only part of the overall process, metadata is sent to the Library of Congress, Bowker, and direct trading partners in order for a title to be discovered on retail sites and in various databases. For many organizations, subprocesses will need to be mapped; the process of creating, approving, proofreading, and archiving catalog copy is in itself a process to examine and improve.
Redundant systems should be avoided whenever possible; limiting the number of different databases in which data is stored helps publishers avoid duplicate (especially manual) entries, minimize data transfer issues, and improve data export capabilities. When multiple systems are utilized, interoperability can be a challenge. Controlling access privileges aims to mitigate inadvertent changes and mistakes; within the organization, not everybody should be able to edit metadata records, but everyone should be able to see or access it. Bob Oeste, a metadata expert and recipient of the 2014 AAUP Constituency Award from the Association of American University Presses for his efforts as metadata guide and evangelist, describes a “Goldilocks Zone” squarely between too much control, or data dictatorship, and anarchy, where everybody does anything (Oeste 2013).

BISG’s certification program validates the accuracy and timeliness of publisher metadata by providing a qualitative and quantitative assessment. Certification includes automatic validation of certain fields, while a panel of industry experts from organizations such as Ingram, Baker & Taylor, and
Bowker manually reviews other fields. A score of 98 on quantitative aspects and 95 on qualitative gives the publisher a Gold rating; a score of 90 on each merits Silver. Some trading partners provide incentives to rated presses; certification feedback gives a publisher clear insight into what aspects need improvement. Six publishers are currently certified gold—four majors and two University Presses (BISG 2014d).

(Reproduced with permission from BISG 2014d)

Until now, BISG’s Product Data Certification Program has been a largely manual effort, but the organization is working on retooling the program to make it more automated and more scalable. Then-Executive Director of BISG, Len Vlahos (2014), mentions that BISG has entered into an arrangement with BookNet Canada, which has a product data certification program called Biblioshare; a joint US version is planned for a beta program at the end of 2014.
To Meta-infinity…and Beyond!

The metadata universe is expanding exponentially as data is becoming increasingly interconnected. OCLC’s WorldCat database informs users which libraries near the user hold a particular title in stock, accessible directly through a Google Books search by clicking “find in a library” from a book’s record. Connected to a university’s IP network, WorldCat shows a user if the book is in the university’s catalog or available in a consortium library; if the user is not on an authenticated network it will refer the user to the closest public library. For example, Cookbook Finder, an OCLC research prototype, provides access to thousands of cookbooks and other works about food and nutrition described in library records; links refer users to related works by author or topic, as well as to full-text versions, when available, from Project Gutenberg or HathiTrust.

MARC Records are the library world’s key metadata standard. Like ONIX, MARC (Machine-Readable Cataloging format) is a descriptive metadata standard for the information industry. Some publishers send MARC Records directly to libraries, though most libraries receive MARC Records from OCLC or other sources. By ensuring that MARC records exist for their titles, publishers can help libraries catalog titles quickly and accurately; this is particularly germane for publishers that sell e-books or electronic subscriptions directly to libraries.

In 2009, OCLC introduced a pilot project to enrich metadata by merging publishers’ ONIX files with MARC and the WorldCat database (Godby 2010, 12–13). The initiative (which I participated in) offered publishers a way to add enhanced data from WorldCat records to ONIX, such as Library of Congress (LC) subject headings missing from the publisher’s own metadata, or to correct certain fields such as author names with authorized name headings. According to Maureen Huss (2014), Director of Contract Services at OCLC, the cooperative
discovered that most publishers were grappling with the new realities of e-books and implementing new pricing models, and the measure to enrich publisher metadata was ahead of its time. OCLC learned lessons from the project that informed internal measures, improving the Functional Requirements for Bibliographic Records (FRBR) work concept, for example, a way to both distinguish between different works and group variations of a single work. A key WorldCat component, FRBR enables the display of e-books, paperbacks, and all sundry varieties of format in separate yet conjoined fashion. Other lessons learned will inform efforts as OCLC plans to reintroduce the program in 2015.

Libraries often face considerable hurdles incorporating e-books from distributors and individual publishers into their catalog systems and integrating metadata from MARC files, ONIX files, and spreadsheets. Metadata “crosswalks” help map and exchange data between two or more metadata schemas, but incorporating sizable e-book purchases into the library system presents a significant task. In particular, these challenges involve dealing with subject headings (translating BISAC into LC subject headings); authorized name headings; special characters; and ensuring receipt of epub files and cover images (Halverstadt and Kall 2013).

Google Scholar, Web of Science, and Scopus are multidisciplinary knowledge databases that have traditionally focused on journals but which now also index books in their databases. Proquest’s Summon Service and EBSCO Discovery Service are popular discovery engines that also include metadata from books, while PubMed Central, LexisNexis, and ERIC are examples of specialized sites that are ubiquitous in certain subject areas. Increasingly, these tools are requesting metadata from publishers, as well as books to use for full text search if not display. Together, these tools enable Linked Data Repositories, which
improve the architecture of the semantic web and, at least potentially, facilitate improved search and discovery (Woodley 2008, 2; Heath and Bizer 2011).

New standards are constantly emerging to improve discovery and accuracy. Published in 2012, the ISNI (International Standard Name Identifier) is an ISO Standard that provides a numerical representation of a name. The ISNI is one of many types of identifiers (BISG has an online interactive guide): ISNIs can be assigned to public figures, content creators or contributors (i.e. researchers, authors, musicians, actors, publishers, research institutions), and subjects of that content (if they are people or institutions, i.e. Harry Potter). ORCID identifiers are used specifically for researchers; ISBNs identify book products; and DOIs (Digital Object Identifier) are used to digitally identify, and locate, “objects” such as journal articles, books, and book chapters, providing a persistent link to where the object, or information about it, can be found on the Internet. The ISNI is a 16 digit unique sequence (example: 0000 0004 1029 5439); the final digit is a check digit and is sometimes an X. In machine-to-machine communication, the ISNI is rendered without the spaces; breaking it up into four sections makes it more human-readable on the web.

The ISNI is a tool for linking data sets and improving disambiguation. For example, the ISNI allows a publisher with two different authors sharing the same name to more readily distinguish between them for royalty purposes. The ISNI helps the publisher remove uncertainty between the two authors, pay correctly, and avoid or reduce the use of sensitive information such as social security numbers.

For discovery, Bowker’s Laura Dawson (2014) uses the example of Brian May. Is the musician the same Brian May who is an astrophysicist, or the Brian May who is the editor of an obscure photography book? The ISNI helps us to determine that Brian May, noted guitarist for the band Queen, is also an
astrophysicist with a PhD who has published his dissertation “A Survey of Radial Velocities in the Zodiacal Dust Cloud,” and also the same person who has annotated a seminal collection of stereoscopic photographs. The ISNI allows for the diverse works of this quite interesting person to be linked together under his unique, persistent identifier.

Brian May: ISNI Record

(Reproduced with permission from Dawson 2014)

The ISNI is also a tool for linking data sets together, such as Books in Print and Musicbrainz, allowing for a music professor to be unambiguously identified in his capacity as a session musician as well as the author of monographs on the evolution of jazz keyboard styles. As Dawson (2014) mentions, “An organization using both of these data sets would be able to link all the work produced by that professor regardless of whether it is audio or text.”
Metadata standards and their implementation are works in progress and e-book metadata is the wild frontier. E-book metadata is inherently complex, due to different formats such as PDF, epub, and mobi or Kindle; digital rights management (DRM); multiple currencies; chapter-level sales; subscription sales; vendor consortia sales to libraries; and the list of variables is only growing. There is little agreement on how to apply ISBNs to different formats—some publishers assign a single ISBN to multiple formats, others assign a different ISBN to each format, and some publishers assign distinct ISBNs to different formats and aspects such as presence of DRM or not, different vendors, and the like. It is still difficult to link print and e-book metadata, although YBP has made great strides with their GOBI database product for library market.

The DOI—developed by the AAP in 1996 and approved as an ISO standard in 2010—offers potential to increase discovery, sales, and citations for scholarly content, though few publishers have yet to exploit its capability beyond the realm of journals. As an identifier, the DOI is unique in that it can identify a work, a product, or a component, such as a chapter, table or chart; it can be assigned to print as well as e-books; it is actionable, linking to whatever resource the publisher wants it to link to; the DOI itself is permanent but the metadata associated with it, as well as the link that it resolves to, can be changed by its owner at any time; finally, it is powerful in that it can link to multiple resources, such as a publisher website, the content itself, book reviews or social media, and the product on retail sites such as Amazon or Powells (Kasdorf 2012).

Despite publishers’ best attempts at “clean” and robust metadata, vendors obtain metadata from many sources—some which have inaccurate data—and use these sources to merge or overwrite publisher metadata. Changes to publisher metadata on retail (such as Amazon) and database (Ingram or Bowker) sites include alterations to the contributor/author field, BISAC categories, age
range, and measurements (Oeste 2013; BISG 2012, 34). “Ninety-five percent of publishers have had the experience of creating their e-books with one set of metadata and seeing an altered set of metadata at the point of sale, online booksellers like Amazon, Barnes & Noble and Apple,” according to Jeremy Greenfield of Digital Book World (2012). This phenomenon is by no means limited to e-book metadata. In my experience—let’s call it Warren’s Law—the author is invariably the one to find errors in the publisher’s metadata, indubitably due to authors’ predilections to checking their Amazon rankings.

Though not reaching the level of interstellar conflict, the alteration of metadata is a concern and source of controversy between publishers, wholesalers, and retailers. Publishers find that some elements of their metadata have been discarded or modified by trading partners and complain that they don’t get feedback from metadata recipients on changes made to publisher metadata; for their part, retailers contend that they habitually receive incomplete or incorrect metadata from a wide cross section of publishers, while questioning the commitment of senders to make the necessary changes to improve accuracy or provide core elements (BISG 2012, 19–23).

Keywords—words and phrases that describe the content or theme of the work, used to supplement the title, subtitle, author name, and description—are one example of a practical, still-evolving application in ONIX metadata. Using keywords in publisher metadata can be useful for discovery on retailers’ websites as well as to enhance search engine optimization (SEO) on a publisher’s website or other websites that feature book products. Keyword usage is fluid, however, as not all metadata recipients support keywords and not under the same parameters. Keyword assignments can be particularly useful when a consumer’s probable search term is new, jargon, distinctive, or quite specific; a consumer is not likely to know the exact title or author of a book; its title differs
from its theme; many different and distinctive terms can describe the topic; or a book includes topics that are outside the BISAC subjects used. Keywords are particularly useful when consumers are likely to use search terms, such as specific characters, places, or series names (i.e. “Game of Thrones,” Targaryen), that may not appear elsewhere in the publisher’s metadata (BISG 2014b, 5–6).

Choosing a set of effective keywords is both art and science. It is important to be specific, avoiding terms that are so general as to not be useful, but not so specific that the intended audience will be unlikely to use the term. Publishers should choose unique, relevant, and consumer-oriented keywords; single words and phrases of 2–5 words are acceptable, and as many synonyms as appropriate can be used. It is best practice to avoid or limit keywords that are already in the title or author fields; “publishers should always use discretion when repeating information within multiple data points,” according to BISG (BISG 2014b, 4). Publishers should prioritize, placing the most important keywords at the beginning of the field; although a 500 character maximum field-length is recommended, not all vendors use the entire field. Semicolons instead of commas are preferred, and it is best practice to avoid spaces following the semicolon (BISG 2014b, 7–8).

Within a publishing organization, the use of keywords warrants planning and consideration. They can be determined book-by-book, ad hoc, or based on a pre-determined, defined taxonomy. Each approach has particular advantages and disadvantages: an open method allows for flexibility but may result in similar titles having unnecessarily distinct keywords, while a structured taxonomy provides standardization but may be rigid or focused less on consumer-oriented search words and more on corporate-approved terminology. Tools that can help identify appropriate keywords include Wordle (copy/paste the text of the summary, introduction, or entire text into Wordle and the most
popular terms will be highlighted in a word cloud); Google’s AdWords Keyword Planner (useful for finding a list of keywords and how they might perform based on historical statistics of search volume or traffic); Google Trends and Correlate; and a variety of SEO keyword research tools.

When Metadata Met Content

A few years ago at the publishing conference Book Expo America, I overheard a colleague mention, “Metadata is the content.” I believe it was (mostly) said in jest, but the statement is not entirely inaccurate.

(Reproduced with permission from Small Demons 2012)
In an (unfortunately failed) experiment in creating metadata from narrative, So-Cal start-up Small Demons (2012) attempted to glean and share the interconnectedness between books and popular culture. Extracting from a book’s contents the music, films, TV shows, and other books mentioned within, Small Demons intended to explore the links between these media and the culture at large. The content became the metadata. Small Demons invited the reader, for example, to browse music mentioned in Jennifer Egan’s *Visit from the Goon Squad*; clicking on Jimi Hendrix’s “Foxy Lady” would transport the reader to other books that featured the song, winding our way down a path of culture’s digital crumbs.

Both an extension and application of metadata, an API (application programming interface) offers programmers an opportunity to retrieve and display metadata and book content in multiple ways, such as O’Reilly’s Product Metadata Interface, which uses RDF (Resource Description Framework) to provide metadata elements on O’Reilly titles. For Kate Pullinger’s new novel, *Landing Gear*, Random House of Canada (2014) created an API in which an excerpt-length section of *Landing Gear* is stored in a content management system and tagged to define characters, locations, events, and times. Programmers can access this data and build new products with it, such as an interactive map, events on a timeline, or character locations during key events. For Pullinger (2014), a digital-savvy author who created (with Chris Joseph) the wonderfully inventive digital novel *Inanimate Alice*, the API provided a new way to experiment beyond the constraints of the page.

Publishers are exploring partnerships to provide hyperlinked resources integrating multimedia, research literature, and reference material, all powered by robust metadata. A partnership between Oxford University Press and Alexander Street Press links OUP’s Oxford Music Online with the latter’s
Classical Music Library, enabling users (at a library that subscribes to both) to jump from sound and bibliographic files in one database to the other (Harley et al. 2010, 615–16).

The Learning Resource Metadata Initiative (LRMI) is a project led by the Association of Educational Publishers (an arm of the Association of American Publishers) and Creative Commons and funded by the Bill & Melinda Gates Foundation. LRMI encourages publishers to tag and describe educational content via a common metadata framework to improve search results and access for online educational resources. Metadata properties that describe educational resources include educational alignment, purpose, age range, and interactivity, among others. LRMI is one of several interrelated efforts to describe and tag educational resources; others include Schema.org, SCORM, and the Learning Registry. LRMI (2014, 3) reports that 58% of surveyed publishers tag their resources with metadata, up from 55% in 2013, and 56% of those who don’t currently tag have plans to begin within the coming year, while 66% of educators surveyed believe publishers should tag educational resources to enable more precise online searching.

Making a distinction between the type of annotation (and hence, metadata) aiding the discovery of online educational materials and that utilized in the support of students, Peter Brantley (2014) declares, “More technical work has been accomplished in data description and discovery then [sic] in elucidating how more evanescent annotation might empower the learning process.”

Receiving data on students’ interactions with learning materials, including digital textbooks, offers the prospect of increasing the quality of textbooks and improving learning, as publishers discover which chapters resonate with students or which aspects cause confusion. Carnegie Mellon’s Open Learning Initiative has collaborated with the NEXUS Research and Policy Center to collect
and analyze data on how students interact with electronic learning materials in order to improve learning and design (Lingenfelter 2012, 20).

The ability of retailers, publishers, and teachers to track students’ reading and interactions with texts, however, raises significant privacy concerns. CourseSmart was founded in 2007 as a joint venture backed by Pearson, Cengage Learning, McGraw-Hill Education, John Wiley & Sons Inc., and Bedford Freeman Worth and then acquired by Ingram in 2014. CourseSmart has tested a program to track and provide teachers with metrics on a student’s “engagement index” as well as students’ activity with highlights, notes, and bookmarks. “It’s Big Brother, sort of, but with a good intent,” according Tracy Hurley, the dean of the College of Business, Texas A&M University (Streitfeld 2014). Adobe’s Digital Editions e-book and PDF app has been reported to send logs of a reader’s activity—including user ID, book metadata, duration of a book read, and even a list of other books on the reader’s app. It sends this information over the Internet, in plain text, to Adobe’s servers “allowing anyone who can monitor network traffic (such as the National Security Agency, Internet service providers and cable companies, or others sharing a public Wi-Fi network) to follow along over readers’ shoulders” (Gallagher 2014).

Data mining and learning analytics can utilize social or “attention” metadata to gauge user engagement, identify students at risk, and improve feedback to both teachers and learners (U.S. Department of Education 2012, 14). This is part of an accelerating trend moving from publisher- or library-created metadata to socially created, or crowd-sourced metadata, to help locate, categorize, and annotate resources.

Metadata about books is not the same thing as metadata about readers of those books, but we can expect these to converge in the future. Integrating “social metadata”—content generated by social media that helps audiences
discover, value, and understand a site’s content—into workflows and community-building efforts is metadata’s vanguard. Tags and other user-generated content enhance social book sites like Goodreads (now owned by Amazon) and LibraryThing, while their APIs are used to integrate social content into other platforms. Meanwhile, publishers supplying metadata to these sites seek to improve the timeliness and accuracy of information about new or backlist titles that users add to their “shelf” and review. Bookmarks and highlights can already be shared across users of devices such as Amazon Kindle. Libraries, archives, and museums employ shared annotations, comments, and tags, seeking to expose collections to a wider audience, augment metadata, and increase patron engagement (Smith-Yoshimura and Holley 2012, 7–13). Hypothes.is (2014) aims to be an open, collaborative platform for sentence-level critique and community peer-review of books, scientific articles, web pages, data, and other knowledge materials. A non-profit organization funded by the Knight, Mellon, Shuttleworth, and Sloan Foundations as well as contributions from individuals, Hypothes.is builds on interoperability framework of the Open Annotation Collaboration project and best practices from the Linked Data effort.

Social Book, developed by Bob Stein at Institute for the Future of the Book, has developed CommentPress in order to turn any document into a conversation, a book into a rendezvous, a monograph into a forum. Stein’s “Taxonomy of Social Reading” (2010) harkens back to the days of early manuscripts that offered plenty of real estate for annotations in the margins. Social Book’s design purposefully integrates shared discussion into the text, through page, paragraph, and even sentence-level annotations, extending the conversation across an entire work and beyond. Social Book also allows community members to upload texts as epub files that can also contain audio and video files, join groups, and form private reading groups, for instance for
classroom discussions on a text. Judiciously curated works such as Open Utopia (2010–2014), edited by sociologist Stephen Duncombe, can enliven and refresh a work approaching its quincentennial.

Social reading platforms will benefit from flexibility that allows users to determine their preferred level of sharing. You might relish the opportunity to see all readers’ comments on the latest fascinating novel, while a Spanish teacher may want only his or her class to share insights about Don Quixote, and I may be eager to benefit from reading an academic monograph about the early Americas that is amply annotated by experts and peers. User-generated comments will form part of the metadata, facilitating discovery of these works, which will in turn form part of Linked Data resources that empower readers to rocket through galaxies of knowledge and creativity.

Metadata is full of paradox. Publishers must exert control and perfect metadata for their titles, while also unleashing it into the world and therefore inviting loss of control and imperfection. Metadata is highly technical yet eminently creative. Even as metadata has become a key component in propelling
discovery and sales of books and other content, it is evolving from the enterprise of illuminating meaning and significance in an individual title to enabling linkages between a broad network of objects and resources. Metadata helps solve queries, explain answers, and distinguish between legitimate and deceptive sources of information in a world in which the volume of information and the complexity of search are increasing inexorably. Combined with statistical text analysis, metadata holds promise to increase the data-rich aspects of social science, offer highly targeted search results, improve the learning effectiveness of digital textbooks, and identify specific thematic areas of commercial promise to publishers. Metadata, connecting producers, distributors, and users of content, weaves together the rich tapestry of information in the digital age.
References

(Note: URLs below as of October 15, 2014, unless otherwise noted.)

Aden, Cindy (Director, Partner Programs at OCLC). 2014. In discussion with author.
https://www.bisg.org/product-data-certification-program-1


Pullinger, Kate. 2014. In discussion with author. See also http://www.katepullinger.com/digital/.


**Acknowledgements**
The author would like to acknowledge Michael Jon Jensen, who reviewed an early draft of this article and provided several valuable suggestions; Cindy Aden, who provided insights into the history of the ONIX standard as well as suggestions that improved the article’s focus; Maureen Huss, who shared information about OCLC’s metadata enrichment project; Len Vlahos, who clarified aspects of BISG best practice guidelines; Bob Oeste, who has imparted generous guidance to the University Press community on metadata issues; and Greg Aden, who has provided the author with generous professional expertise on metadata and digital publishing. Any errors or omissions in this article are the author’s own.
Author Bio

**John W. Warren** is Head of Mason Publishing Group/George Mason University Press and an adjunct professor in George Washington University’s master of professional studies in publishing program. Warren was previously Marketing and Sales Director at Georgetown University Press, and Director of Marketing, Publications, at the RAND Corporation. He has a Masters in International Management from the Graduate School of International Relations and Pacific Studies (IR/PS) at the University of California, San Diego. He has delivered keynote presentations at several international conferences and has authored several articles about the evolution of e-books, including “Innovation and the Future of E-Books” (2009), for which he was the winner of the International Award for Excellence in the development of the book.