Zen and the Art of Metadata Maintenance

John W. Warren
Head, Mason Publishing Group & George Mason University Press
May 29, 2015
The magnitudes of metadata
Metadata and the Powers of Ten

• In the 1968 film *Powers of Ten* by Charles and Ray Eames, we start with a picnic by the lakeside in Chicago, and are transported every ten seconds by a magnitude of ten to the outer edges of the universe. Returning to earth, we move into the hand of the sleeping picnicker—and with ten times more magnification every ten seconds we end inside a proton of a carbon atom within a DNA molecule in a white blood cell.

• Metadata is like that: it is both a universe and DNA.
Metadata—“data about data”—is the lifeblood of publishing in the digital age and the key to discovery
Metadata is a universe of standards across the information spectrum

From “Seeing Standards: A Visualization of the Metadata Universe”
Communities include Archives, Museums, and Libraries, among others.
Zooming into the domain of scholarly texts we find a plethora of standards, including ONIX.
The goal of good book metadata is to make your book discoverable...and ultimately to be purchased.
Metadata begins with basic information

Product basics: Title, author, and jacket cover
Product description: Catalog copy, backcover copy
Product details: including identifiers (ISBN), page counts, price(s), carton quantity
Additional information: author bios, blurbs, reviews, awards received, table of contents, keywords, audience code, rights information, other currencies
And even more information: book trailers, sample chapters, excerpts, tags, other books, regional info
Metadata is integral to the entire publishing ecosystem

• Book metadata commonly thought of as descriptive data or product information (title, author, etc.)
• Can be defined broadly as information that describes content, or an item that describes another item
• In its broader sense, a book review becomes part of a book’s metadata, while the book review itself is content with its own metadata
• A book becomes part of the metadata of each book or article it references, the author becomes part of the book’s metadata and the book part of the author’s metadata
Metadata is information flow

Large Publishers

Medium-size Publishers

Smaller Publishers

Metadata management services (not always used)

Distributors and aggregators (e.g. wholesalers, data companies, etc.)

Retailers (e.g., B&N, Amazon, indie stores, etc.)
Metadata is messy

Large Publishers

Medium-size Publishers

Smaller Publishers

Metadata management services (not always used)

Distributors and aggregators (e.g. wholesalers, data companies, etc.)

Content converters

Retailers (e.g., B&N, Amazon, indie stores, etc.)

Other Consumer Facing Companies (e.g., social sites, etc.)
Metadata, though not called by that name, has been important to booksellers, libraries, and collectors nearly as long as the book itself.
Metadata is clearly on the rise
ONIX is a standard for representing book, serial and video product information in electronic form

<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>
</Product>
ONIX—ONline Information Exchange—has evolved and is evolving

- ONIX was developed and is maintained by EDItEUR, jointly with Book Industry Communication (UK) and Book Industry Study Group (US), in cooperation of industry stakeholders
  - ONIX 1.0 released January 2000
  - ONIX 2.1 released 2004
  - ONIX 3.0 released 2009
  - ONIX 3.01 released January 2012

- Many industry partners are still not accepting ONIX 3, although EDItEUR/BISG plan to sunset ONIX 2.1

- User groups in 13 different countries

- Best practices guides and code specifications available on BISG and EDItEUR websites
ONIX feeds are sent to book industry trading partners

- 32 “core” metadata elements; 27 of these required
- Enriched ONIX data can include hundreds of fields of data such as description, rights information, reviews

Trading partners include:
- **Wholesalers** (Ingram, Baker & Taylor)
- **Online retailers** (Amazon, Powells)
- **Retail chains** (Barnes and Noble)
- **Databases** (Library of Congress, Bowker Books in Print, Serials Solutions, OCLC/WorldCat)
Using the ONIX code to indicate your audience is one example

**ONIX Code List 28 Audience Code List** (Audience Code, Description)

- 01 - General/Trade
- 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
- 08 - Adult Education
ONIX is preferred, but most trading partners will take other methods

- Other common formats include excel or comma delimited file formats
- Some trading partners will NOT accept ONIX
- Some vendors have “special needs” —be prepared to accommodate important trading partners
- Cover files are also sent in a variety of formats/sizes
- According to BISG, subject code is the most common error in publisher metadata
- Publishers, for their part, also claim that BISAC subject headings are the most frequent element modified in the supply chain by recipients
Metadata matters

Metadata: average sales per ISBN

- Incomplete BIC, no image
- Incomplete BIC, with image
- Complete BIC, no image
- Complete BIC, with image
Metadata is a process of continuous improvement

• Improvement efforts should include increasing the **quality** of metadata as well as the **quantity** of accounts to which you feed metadata directly in ONIX
• Effort never ends—there’s always something to add, improve, clean up
• Data constantly changes; errors are introduced regularly
• Limit the number of different databases you store data
• Not everybody should be able to edit it, but everyone should be able to see it/access it
• Determine the “Goldilocks Zone” between control/dictatorship and anarchy/everybody does anything
• Mapping your metadata process is a good way to start
Metadata flows through numerous systems to trading partners, and ultimately to thousands of destinations.
Metadata is sent to Library of Congress, Bowker, and direct trading partners...
... In order for your titles be discovered on retail sites and in various databases.
Metadata holds a central role in the publishing organization

- **Administration**: 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 or micro segments
- **Learning**: integration of content into learning management systems and databases, from Blackboard to SCORM
- **Preservation**: digital assets management, version control
- **Use**: sales tracking, search logs, and ebook usage
BISG certification validates your metadata’s accuracy and timeliness

- Qualitative and quantitative look at publisher data
- Automatic validation of certain fields
- Panel of industry experts manually reviews other fields
- Score of 95 quantitative / 90 on qualitative gives Silver rating
  - Ingram gives preferential treatment to rated presses
- BISG working on retooling program to be scalable
- Feedback gives you clear idea of what to improve
Six publishers are currently certified gold — four big ones and two University Presses.
To meta-infinity... and beyond!
OCLC’s WorldCat shows what libraries near the user hold your title in stock.
MARC Records are the library world’s key metadata format

- MARC Records (library standard bibliographic data) can be sent directly to libraries, though most libraries get MARC Records from OCLC or other sources
- MARC Records help libraries catalog titles quickly and accurately
- OCLC has an initiative to merge ONIX and WorldCat/MARC Records
- Metadata “crosswalks” map and exchange data between two metadata schemas
Metadata is becoming increasingly interconnected

- Google Scholar, Web of Science, and Scopus: multidisciplinary knowledge databases
  - Focus on journals — also indexes books in database
- Serials Solutions Summon Service and EBSCO Discovery Service: popular discovery engines
- PubMed Central, LexisNexis, ERIC: specialized sites ubiquitous in certain subject areas
- These tools enable Linked Data Repositories, improving architecture of the semantic web
Metadata can be supplied to social sites such as Goodreads and LibraryThing.
New standards emerge constantly, improving discovery and accuracy

What is the ISNI?

- **International Standard Name Identifier**
- Numerical representation of a name
  - Assigned to public figures, content 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)
  - 16 digits
  - Example: 0000 0004 1029 5439
ISNI is a tool for linking data sets and improving disambiguation.

Brian May, guitarist in Queen, is an astrophysicist, AND is the editor of an obscure photography book (yes, an interesting fellow!)
Researchers (and organizations) can sign up for ORCID IDs

- ORCID provides a persistent digital identifier that distinguishes researchers from others (i.e. of the same or similar name)
- Integrated into key research workflows such as manuscript and grant submission
- Supports automated linkages between researchers and their professional activities, ensuring that their work is recognized
DOI can improve discovery, sales, and citation of scholarly content

- DOI—Digital Object Identifier—developed by the AAP in 1996 and approved as an ISO standard in 2010
- Few publishers have used DOI beyond realm of journals
- Can identify a work, product, component of a work such as a chapter, table, or chart
- DOI is actionable, links to whatever resource the publisher wants it to link to
- DOI itself is permanent but its metadata, as well as the link that it resolves to, can be changed by its owner at any time
- Can link to multiple resources: publisher website, content itself, book reviews or social media, and product on retail sites such as Amazon or Powells
Metadata is a work in progress

- Despite your best attempts at “clean” and robust metadata, vendors get metadata from many sources—some which have inaccurate data
- Ebook metadata is still the wild frontier
- Still difficult to link print and ebook metadata; YBP is good at this for library market
- Increasing complexities of multiple formats, multiple currencies, chapter-level sales, etc.
“95% 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.”
— Jeremy Greenfield, Editorial Director, Digital Book World
Keywords are a practical and still evolving application in ONIX

- Keywords are words and phrases that describe the content or theme of the work that are:
  - Relevant to the work
  - Used to supplement the title, subtitle, author name, description

- Keywords are not required but can be useful for discovery and Search Engine Optimization (SEO)

- Use of keywords is fluid; not all data recipients support use but many utilize at least some keywords

- Publishers should use much discretion before repeating information within multiple data points
Use of keywords can improve discoverability

• Useful when consumer’s search term is jargon, new, distinctive, specific
• When consumer does not know exact title or author of book, or title differs from theme
• When many different and distinctive terms describe the topic
• When book includes topics that are outside the BISAC subjects used or consumer use terms outside metadata
Choosing set of effective keywords is an art and a science

• Be **specific**, but not so specific people won’t use the term
• Choose **unique** and relevant keywords
• **Avoid** terms already in title or author fields
• Choose **consumer-oriented** keywords
• Single words and phrases of 2–5 words are acceptable
• Use as many synonyms as appropriate
• **Prioritize**—500 character max recommended, but not all vendors use the entire field
• Semicolons instead of commas are preferred, and best with no spaces
Title: A Dance with Dragons
Author: George R. R. Martin
ISBN: 9780553801477
Series: A Song of Ice and Fire (Book 5) BISAC Subject Headings: FIC009020; FIC002000; FIC028010

Keywords: Game of Thrones; kingdoms; kings; magic; dragons; HBO series; medieval; saga; Targaryen

Description: In the aftermath of a colossal battle, the future of the Seven Kingdoms hangs in the balance—beset by newly emerging threats from every direction. In the east, Daenerys Targaryen, the last scion of House Targaryen, rules with her three dragons as queen of a city built on dust and death. But Daenerys has thousands of enemies, and many have set out to find her. As they gather, one young man embarks upon his own quest for the queen, with an entirely different goal in mind.
Overheard at Book Expo America

(a few years ago)

“Metadata is the content”
Social metadata is on the rise
Metadata can even be a game—fun!

In our first collaborative game with the British Library, tag naval images from the library's immense collections!

Wished there was a fun way to organize and understand the world's archived books? Check out Book Tag!
Social reading is metadata’s next frontier

• Shared and user-created bookmarks, tags, highlights, annotations, comments, notes

• **Social Book**, developed by Bob Stein at Institute for the Future of the Book, “turns each document into a conversation”

• **Hypothes.is** is an open, collaborative platform for critique and community peer-review of books, scientific articles, web pages, data, and other knowledge materials

• Publishers interested in receiving data on ebook reading habits and student interaction with etextbooks
  – Potential to increase quality of textbooks and perhaps improve learning
  – Privacy concerns are not insignificant
When metadata met content
The novel becomes an API

- For *Landing Gear*, Kate Pullinger’s new novel, Random House created an API (application programming interface) that allows programmers to get content in multiple ways

- Excerpt-length section of *Landing Gear* stored in a content management system is tagged to define characters, locations, events and times

- Programmers can access this data and build new products with it, such as plot locations on a map or events on a timeline, or even character locations during key events
In this example, the API shows you what happens to Yacub at the Karachi airport.

Yacub - Karachi Airport

"At the airport, I followed the instructions Ameer had given me—for which I had paid the last of my Dubai money—and found the unlocked door that led outside to the planes. I had less than fifteen minutes after darkness fell to find the correct..."

"From the ground, the planes looked enormous, their lights blinking in the dusk. The air..."
Discussion/Questions

• Please contact me with any questions:
  
  John W. Warren  
  Head, Mason Publishing Group/George Mason University Press, George Mason University  
  Email: jwarre13@gmu.edu  
  Twitter: @john_w_warren

• Based on the article: “Zen and the Art of Metadata Maintenance,” John W. Warren, June 17, 2015, Journal of Electronic Publishing  
  http://dx.doi.org/10.3998/3336451.0018.305
Sources and Further Reading (1)

BISG: [https://www.bisg.org/publications](https://www.bisg.org/publications)
- ONIX: [https://www.bisg.org/onix-books](https://www.bisg.org/onix-books)

CrossRef: DOI for books: [http://www.crossref.org/02publishers/dois_for_books.html](http://www.crossref.org/02publishers/dois_for_books.html)


Hypothes.is: [https://hypothes.is/](https://hypothes.is/)

ISNI database: [http://www.isni.org/search](http://www.isni.org/search)

Sources and Further Reading (2)

Metadata games: http://www.metadatagames.org/

ORCID: http://orcid.org/

Open Utopia on Social Book: http://theopenutopia.org/social-book/


Uncharted: Big Data as a Lens on Human Culture, Aiden, Erez and Jean-Baptiste Michel, 2013, New York: Riverhead