Society for Scholarly Publishing Conference Arlington, VA

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

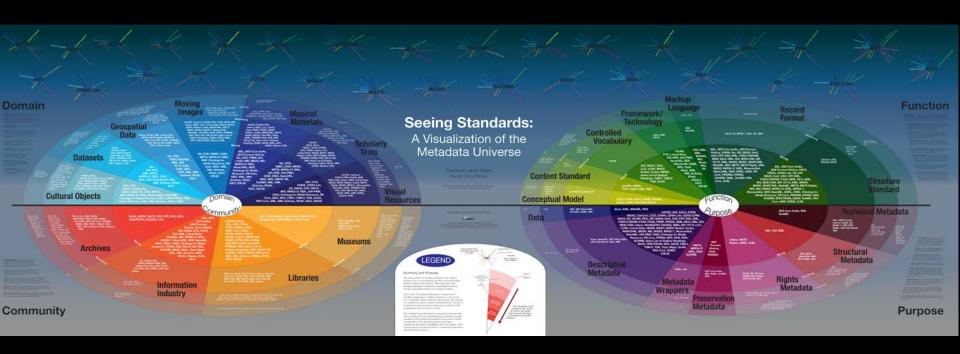


Metadata and the Powers of Ten

- 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

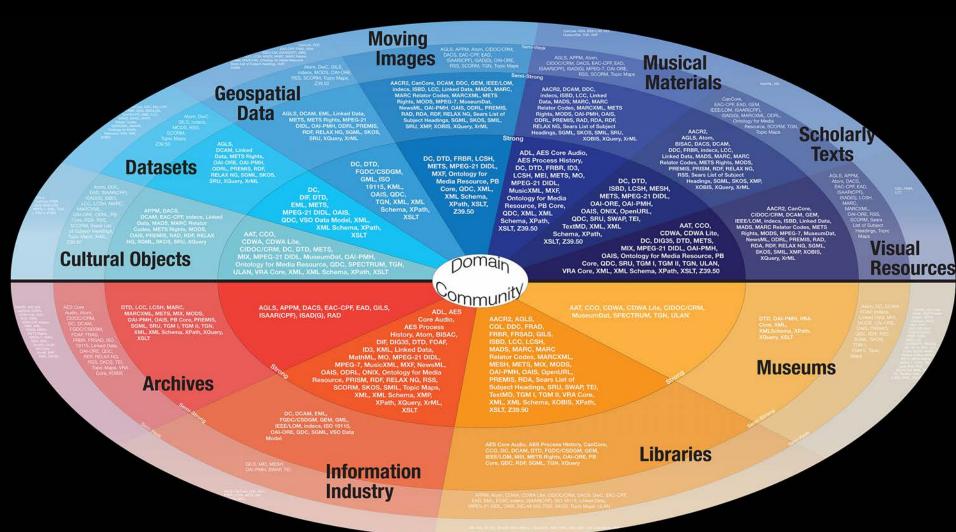
Metadata—"data about data"—is the lifeblood of publishing in the digital age and the key to discovery

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

CIDUC/CHM, DACS, EAC-CPF, EAD, Musical pic Maps ISAAR(CPF), ISAD(G), MPEG-7, OAI-ORE, RSS, SCORM, Topic Maps Semi-Strong **Materials** LOM, ARC. AACR2, DCAM, DDC. indecs, ISBD, LCC, Linked CanCore. Data, MADS, MARC, MARC EAC-CPF, EAD, GEM, Relator Codes, MARCXML, METS IEEE/LOM, ISAAR(CPF), Rights, MODS, OAI-PMH, OAIS, of ISAD(G), MARCXML, ODRL, ODRL, PREMIS, RAD, RDA, RDF, Ontology for Media **RELAX NG, Sears List of Subject** Resource, SCORM, TGN, AACR2, Headings, SGML, SKOS, SMIL, SRU, Topic Maps Strong AGLS, Atom, XOBIS, XQuery, XrML BISAC, DACS, DCAM, DDC, FRBR, indecs, LCC, ADL, AES Core Audio. Linked Data, MADS, MARC, MARC **AES Process History.** Relator Codes, METS Rights, MODS, DC, DTD, FRBR, ID3, PREMIS, PRISM, RDF, RELAX NG, LCSH, MEI, METS, MO. RSS. Sears List of Subject DC, DTD, Headings, SGML, SKOS, XMP, MPEG-21 DIDL. XOBIS, XQuery, XrML ISBD, LCSH, MESH, MusicXML, MXF, METS, MPEG-21 DIDL, **Ontology for Media** OAI-ORE, OAI-PMH, Resource, PB Core, AACR2, CanCore, OAIS, ONIX, OpenURL, QDC, XML, XML CIDOC/CRM, DCAM, GEM, QDC, SRU, SWAP, TEI, Schema, XPath, IEEE/LOM, indecs, ISBD, Linked Data, AAT, CCO. TextMD, XML, XML MADS, MARC Relator Codes, METS **XSLT, Z39.50** CDWA, CDWA Lite, Schema, XPath, Rights, MODS, MPEG-7, MuseumDat, DC, DIG35, DTD, METS. NewsML, ODRL, PREMIS, RAD, XSLT, Z39.50 RDA, RDF, RELAX NG, SGML, MIX, MPEG-21 DIDL, OAI-PMH, SKOS, SMIL, XMP, XOBIS,

OAIS, Ontology for Media Resource, PB

Core, QDC, SRU, TGM I, TGM II, TGN, ULAN,

VRA Core, XML, XML Schema, XPath, XSLT, Z39.50

Scholarly Texts

XQuery, XrML

AGLS, APPM,
Atom, DACS,
EAC-CPF, EAD,
ISAAR(CPF),
ISAD(G), LCSH,
MARC,
MARCXML,
OAI-ORE, RSS,
SCORM, Sears
List of Subject
Headings, Topic
Maps

Vis Res

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

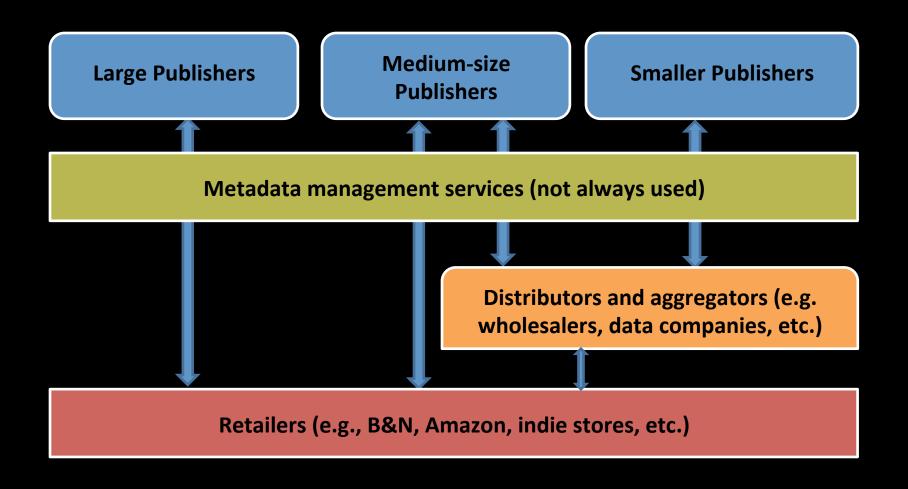
Additional information: author bios, blurbs, reviews, awards received, table of contents, keywords, audience code, rights information, other currencies

And even more information: book trailers, sample chapters, excerpts, tags, other books, regional info

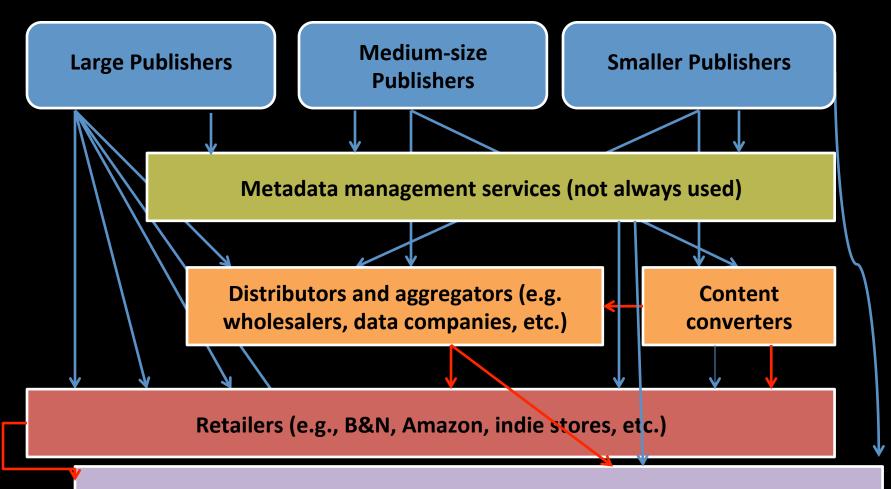
Metadata is integral to the entire publishing ecosystem

- 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

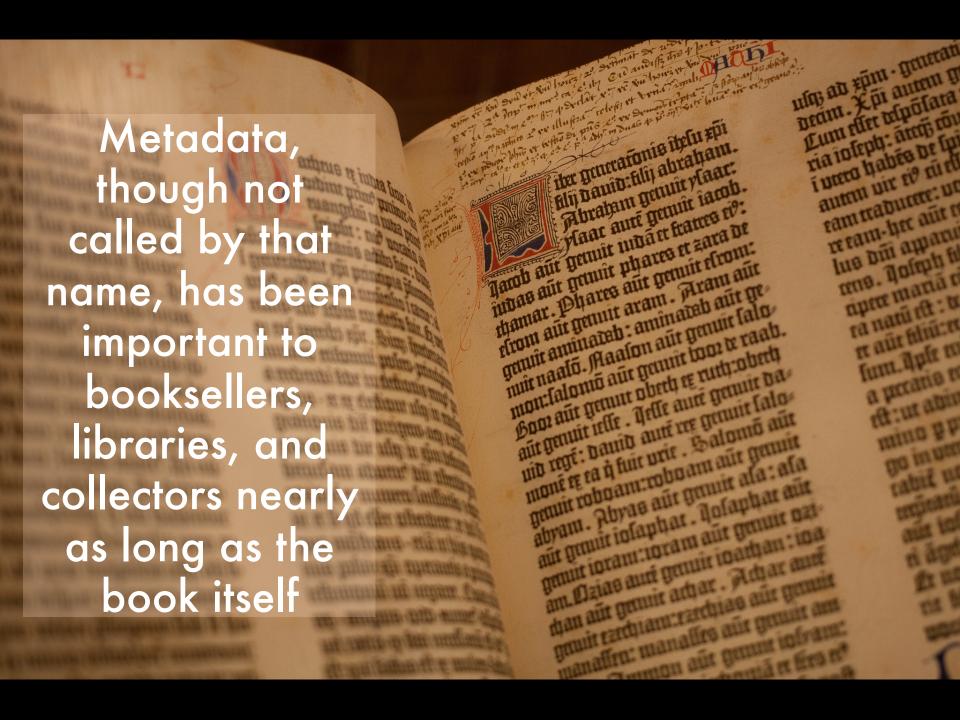
Metadata is information flow



Metadata is messy

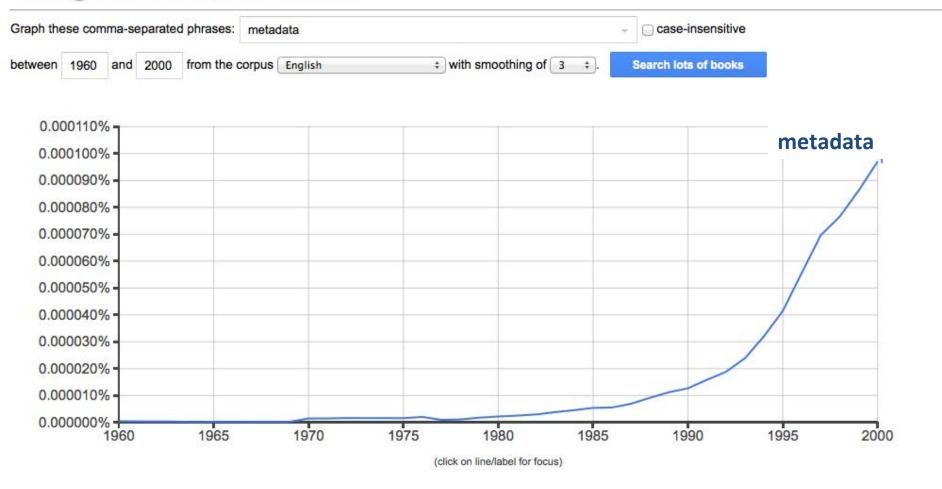


Other Consumer Facing Companies (e.g., social sites, etc.)



Metadata is clearly on the rise

Google books Ngram Viewer



ONIX is a standard for representing book, serial and video product information in electronic form

```
<Product>
       <RecordReference>1234567890</RecordReference>
       <Notification Type>03</NotificationType>
       <ISBN>9781626160132</ISBN>
       <Pre><Pre>oduct Form>BB</Pre>
       <DistinctiveTitle>Hezbollah</DistinctiveTitle>
       <Contributor>
               <ContributorRole>A0|</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—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 I.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

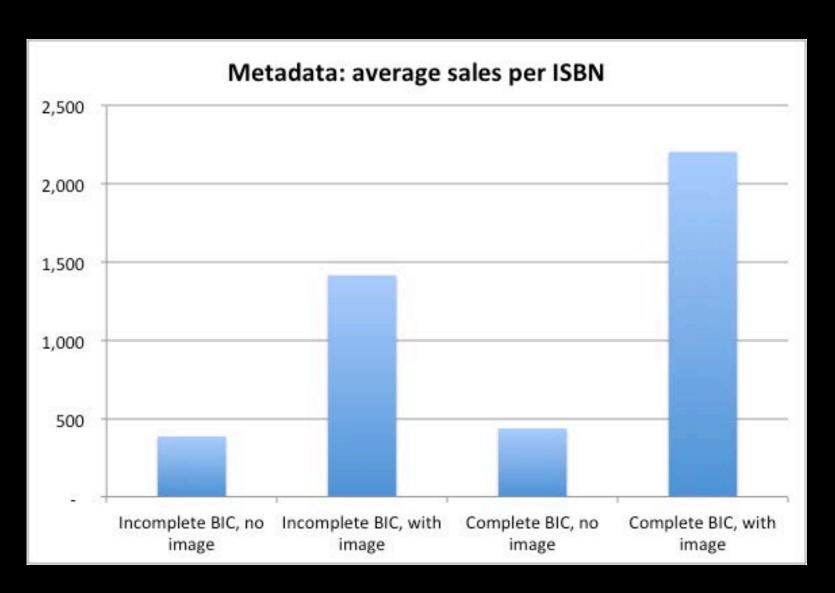
ONIX Code List 28 Audience Code List (Audience Code, Description)

- 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

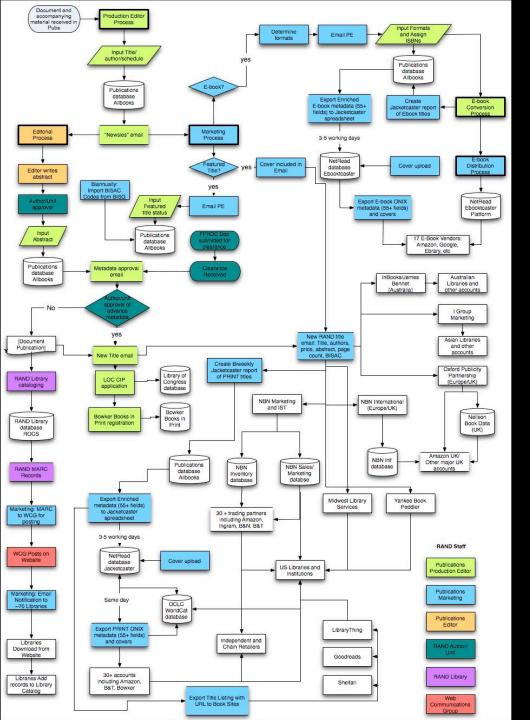
- Other common formats include excel or comma delimited file formats
- Some trading partners will NOT accept ONIX
- Some vendors have "special needs" —be prepared to accommodate important trading partners
- Cover files are also sent in a variety of formats/sizes
- According to BISG, subject code is the 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



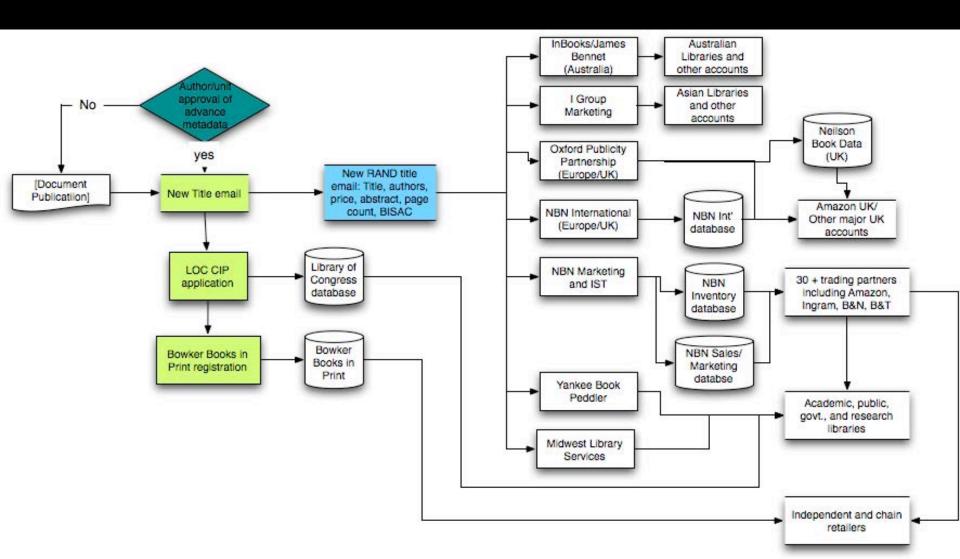
Metadata is a process of continuous improvement

- 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



The author, a retired CIA officer, provides insights into these events from an intelligence professional's perspective, highlighting the tradecraft of intelligence collection, counterintelligence, and covert actions

and relating how many of the principles of the era's intelligence practice are still relevant today. Daigler

Read more

Add to Wish List



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

PDCP: Currently Certified Companies

The following organizations have been successfully certified through BISG's Product Data Certification Program. Click here to read about the different certification levels.

GOLD CERTIFIED COMPANIES









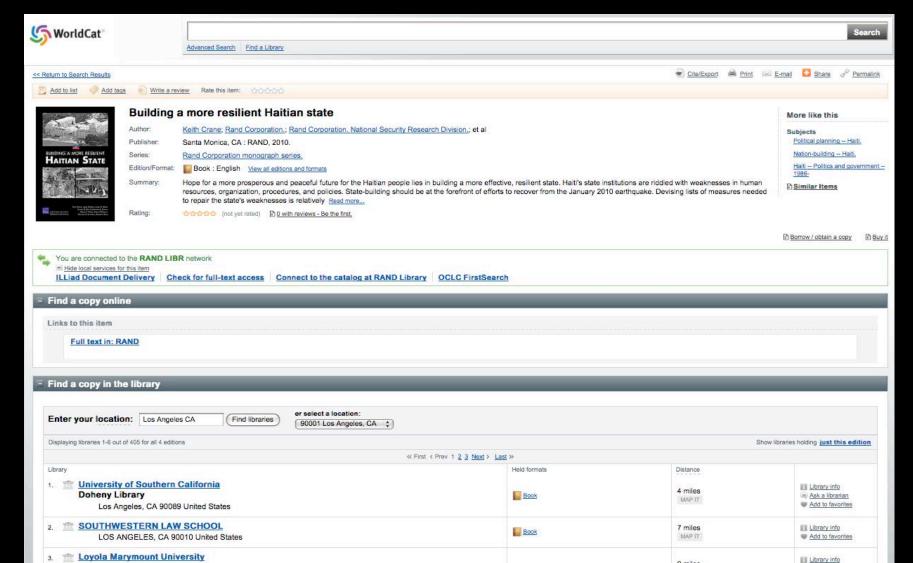


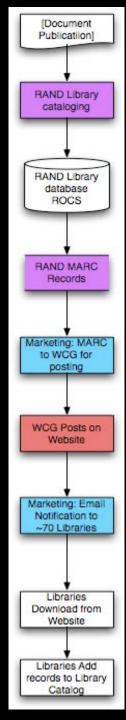


Six publishers are currently certified gold -four big ones and two University Presses

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?

- ISO Standard, published in 2012
- 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

Show Must Go On, The Sleeping on the Sidewalk Slow down Some Day One Day Somebody to love . "Alguien a quien amar" Son and Daughter Star fleet Star fleet (8 min) Star fleet project Star fleet project [SR] p1983: Stars & guitars: the guitars that made 200 rock Gods famous Strat pack live in concert, The survey of radial velocities in the zodiacal dust cloud, A Sweet Lady Teo Torriatte (Let Us Cling Together) Tie Your Mother Down Too much love kill you Too Much Love Will Kill You Too much love will kill you (4 min 29 s) Under Pressure village lost and found, A: an annotated tour of the 1850s series of stereo photographs "Scences in our villate" by T. R. Williams village lost and found, A: an complete annotated collection of the original 1850s stereoscopic photograph series "Scences in our village" by T. R. Williams want to break free. I We are the champions We Will Rock You

We will rock you : recorded by Queen We will rock you (3 min 39 s)live version 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

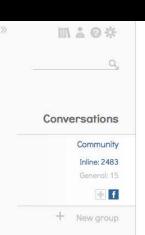
Description: In the aftermath of a colossal battle, the future of the Seven Kingdoms hangs in the balance—beset by newly emerging threats from every direction. In the east, Daenerys Targaryen, the last scion of House Targaryen, rules with her three dragons as queen of a city built on dust and death. 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



[OPEN] UTOPIA, THOMAS MORE

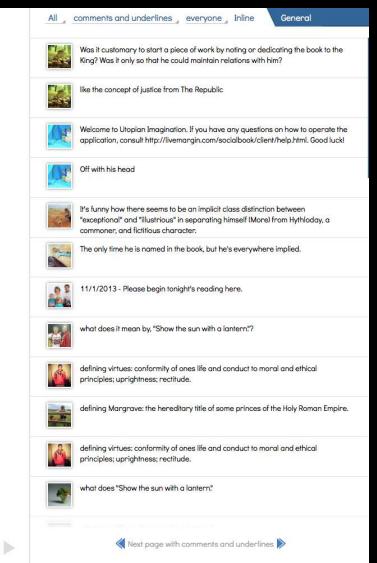
Book I

FIRST BOOK
OF THE DISCUSSION WHICH THE EXCEPTIONAL MAN
RAPHAEL HYTHLODAY HELD CONCERNING
THE BEST STATE OF A COMMONWEALTH,
BY WAY OF THE ILLUSTRIOUS MAN THOMAS MORE,
CITIZEN AND UNDERSHERIFF OF THE
GLORIOUS CITY OF LONDON IN BRITAIN

Henry VIII, the unconquered King of England, a prince adorned with all the virtues that become a great monarch, having some differences of no small consequence with Charles the most serene Prince of Castile, sent me into Flanders, as his ambassador, for treating and composing matters between them. I was colleague and companion to that incomparable man Cuthbert Tunstall. whom the King, Cuthbert Tunstall. with such universal applause, lately made Master of the Rolls; but of whom I will say nothing; not because I fear that the testimony of a friend will be suspected, but rather because his learning and virtues are too great for me to do them justice, and so well known, that they need not my commendations, unless I would, according to the proverb, "Show the sun with a lantern."

Those that were appointed by the Prince to treat with us, met us at Bruges, according to agreement; they were all worthy men. The Margrave of Bruges was their head, and the chief man among them; but he that was esteemed the wisest, and that spoke for the rest, was George Temse, the Provost of Casselsee: both art and nature had concurred to make him eloquent: he was very learned in the law; and, as he had a great capacity, so, by a long practice in affairs, he was very dexterous at unraveling them. After we had several times met, without coming to an agreement, they went to Brussels for some days, to know the Prince's pleasure; and, since our business would admit it, I went to Antwerp.

While I was there, among many that visited me, there was one that was more acceptable to me than any other, Peter Giles, born at Antwerp, who is a man of great honor, and of a good rank in his town, though less than he deserves; for I do not know if there be anywhere to be found a more learned and a better bred young man; for as he is both a very worthy and a very knowing person, so he is so civil to all men, so particularly kind to his friends, and so full of candor and affection, that there is not, perhaps, above one or two anywhere to be found, that is in all respects so perfect a friend: he is extraordinarily modest, there is no artifice in him, and yet no man has more of a prudent

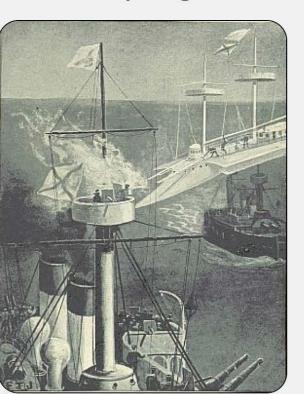


METADATAGAMES

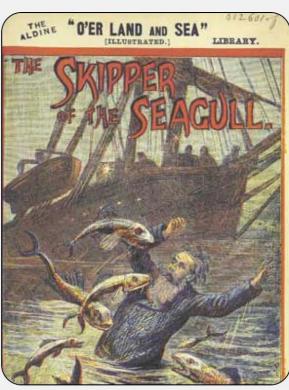
About Y

Get Involved >

Ships Tag



Book Tag



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

Wished there was a fun way to organize and understand the world's archived books? Check out Book Tag!

Metadata can even be a game -fun!

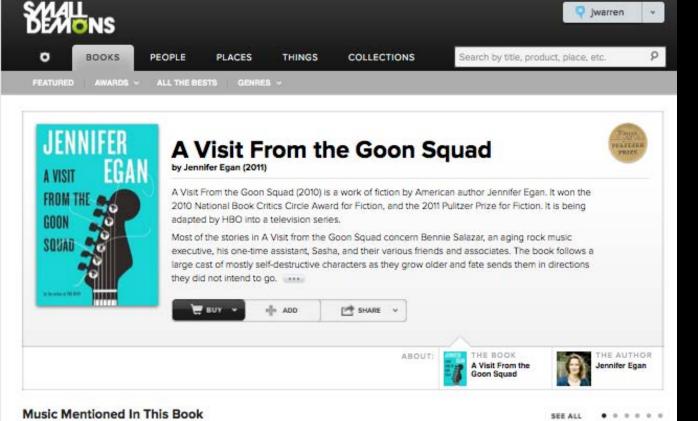
PLAY

immense collections!

PLAY

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



Love the music you see

Download on Tunes

Listen on Spotify

here?

When metadata met

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



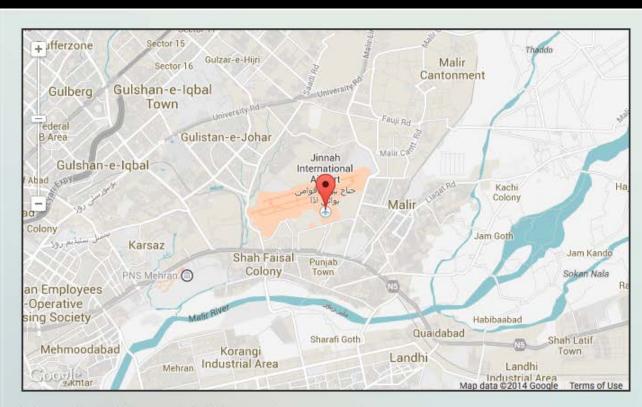
Yacub

Village Dubai Work Camp Karachi

Karachi Airport

Super Market

Corner Shop



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: jwarre | 3@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

Best Practices for Metadata: https://www.bisg.org/product-metadata-best-practices

ONIX: https://www.bisg.org/onix-books

Product Data Certification Program:

https://www.bisg.org/product-data-certification-program-l

Best Practices for Keywords for Metadata

https://www.bisg.org/best-practices-keywords-metadata

CrossRef: DOI for books: http://www.crossref.org/02publishers/dois for books.html

EDItEUR, Codelists Issue 28 for Release 2.1 and 3.0:

http://www.editeur.org/14/Code-Lists

Hypothes.is: https://hypothes.is/

ISNI database: http://www.isni.org/search

Linked Data: Evolving the Web into a Global Data Space, Heath, Tom and Christian Bizer,

Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan &

Claypool, 2011: http://linkeddatabook.com/editions/1.0/

Sources and Further Reading (2)

Metadatagames: http://www.metadatagames.org/

ORCID: http://orcid.org/

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

Seeing Standards: A Visualization of the Metadata Universe, Riley, Jenn and Devin Becker. 2009-–2010: http://www.dlib.indiana.edu/~jenlrile/metadatamap/

"Taxonomy of Social Reading: A Proposal," Bob Stein, Institute for the Future of the Book, 2010: http://futureofthebook.org/social-reading/.

"Teacher Knows if You've Done the E-Reading," David Streitfeld, New York Times, April 8 2013:

http://www.nytimes.com/2013/04/09/technology/coursesmart-e-textbooks-trackstudents-progress-for-teachers.html

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

"White Paper: The Link Between Metadata and Sales," Nielsen Book, Prepared by Andre Breedt and David Walter, 2012:

http://www.nielsenbookdata.co.uk/controller.php?page=1129