Are we done solving metadata problems yet?
Presenters

Rice Majors
Associate University Librarian for Collections & Scholarly Communication, Santa Clara University Library

Tim Coates
Library Advisor, Book Retailer and Consultant to Innovative Interfaces

Graham Bell
Executive Director, EDItEUR

Jeff Penka
VP Product, Zepheira
Rice Majors

Associate University Librarian for Collections & Scholarly Communication, Santa Clara University Library
What are we trying to do?

• Why do we need metadata?
• When do we need better metadata and when do we need better systems?
Libraries use metadata

• To support our patrons
  • Searching for known items
  • Searching for known items with an “incomplete” citation
  • Searching & browsing by topic – wanting only what the library “has”
  • Searching & browsing by topic – wanting to read expansively regardless of ownership

• To support library staff & library activities
  • Analyze collection by topic (e.g. for accreditation)
  • Analyze collection by usage [by topic] (e.g. for collection development or collection management purposes like weeding)
What library catalogs can do (well)

- Reflect our inventory
  - Owned print (i.e. tangible) resources

- Reflect inventory that is immediately accessible
  - Electronic resources that are leased or owned
  - Electronic resources that trigger a rental charge (opaque to user)
  - Non-inventory like Open Access resources and other free internet resources

- Reflect inventory that’s already on-order

- Reflect inventory that we can get non-immediately
  - Consortial inventory or inventory at remote locations
  - Print DDA programs where an order is placed in real-time
What library catalogs don’t do (well)

• Reflect inventory that may become available
  • Reflect resources that could be purchased/rented once they are published
• Connect users to interlibrary borrowing
• Connect users to the whole ecosystem of scholarship
  • (Mostly users are kept within the domain of owned/subscribed)
• Let users fix errors in metadata or add metadata
• Tell users where in the physical collection to go browse
  • Users infer this rather than being explicitly given the information
What have we believed we needed?

• Consistency in metadata quality
  • If some resources are better-described than others, users may believe they have found “everything” when they do a search and get partial results
  • YET: catalogs with AACR1-to-present records living together; fiction without 6xx fields; the problem(s) of cataloging records for videos

• High quality metadata
  • Users cannot find records that do not have subject analysis, etc.
  • YET: catalogs with thousands of vendor-supplied records for e-resources that lack these exact things

• (Sometimes) locally tailored metadata
SEE ALSO: The rest of the world

• Patrons have been trained to try Google (and Amazon, etc.)
• Google is trying to connect users to the entire open web
  • (Many) users want to access the entire ecosystem of scholarship
  • Library catalogs try to keep users in a closed system, even underpromoting ILL
• Google, Amazon offer some semantic searching support
  • Libraries only load authority data for materials they actually own
  • Library systems leverage authority data less aggressively
• Google, Amazon still sometimes fail at guessing
  • What do you do if your result isn’t on the first page?
Potential improvements

• View your catalog as primarily part of an ecosystem
  • Be more inclusive of resources not immediately available

• Allow for greater flow of metadata between systems with fewer professional human interventions
  • Accept records that are “good enough”
  • Build systems and community practices that allow for easier remediation of metadata by more kinds of people

• Build systems that embrace modern expectations
  • Greater semantic searching support
  • Leverage patron data to suggest resources reactively and proactively
Tim Coates

Book retailer - MD, Waterstone’s, UK and WH Smith Europe
Library supply - General Manager YBP (UK)
Library adviser
Book Supply Chain analyst; Print and Digital
Author, Editor
Bringing libraries and publishers together in a new closer relationship with a common aim
Readers want instant access to everything ...
If libraries don’t provide instant access

The readers will get their material somewhere else
UK public libraries: annual use per person
(source CIPFA: Data from local authorities)

- Book Loans
- Visits
Collection management is not just for libraries

It is also for readers
“Metadata” is not just description: it is the language of the supply chain

Good metadata speeds access
Publishers don’t know what is held, circulated, used or searched for in libraries

They need to, in order to provide fast, efficient, tailored access at the best price
Libraries and publishers need one common metadata language.

At present the publishers use ONIX. It works well.
Libraries need to share information openly with each other, with readers, with authors, with funders and with publishers.
Libraries need to change what they use an ILMS for.

The ILMS should be used to provide readers with instant access to everything; and information to everybody.
About EDItEUR

• a not-for-profit membership organisation
• develops, supports and promotes metadata and identification standards for the global book, e-book and serials supply chains
• acknowledged centre of expertise on standards and metadata for the industry
• based in London, but a global membership of publishers, distributors, wholesalers, retailers, subscription agents, libraries, system vendors, rights organizations and trade associations
ONIX for the book trade

• highly-structured XML-based standard for bibliographic metadata
used by publishers, retailers, distributors, data aggregators in many
countries
ONIX for the book trade

• highly-structured XML-based standard for bibliographic metadata used by publishers, retailers, distributors, data aggregators in many countries
  • subject and audience classification
  • marketing collateral – descriptive text, review snippets, cover, extracts from content, author biog...
  • OA and licensing information
  • accessibility guidance
  • sales rights and supply sources
  • pricing (which can be... complicated)
ONIX for the book trade

• highly-structured XML-based standard for bibliographic metadata used by publishers, retailers, distributors, data aggregators in many countries
  • ONIX data is *dynamic* – it’s distributed several months in advance of publication
  • and updated frequently... even post-publication
  • by thousands of publishers across many countries
  • publishers’ motives are commercial, so authority control isn’t their top priority
ONIX for the book trade

- taken from a Nielsen study of the top-selling 100,000 ISBNs in the UK in 2015/16. A similar Nielsen study for the US shows equivalent findings
...and MARC is for libraries

• but in fact, MARC records and CIP data are already created pre-publication using ONIX data
  • then enhanced later from book-in-hand
  • library and commercial syntactically very different, semantically mostly compatible. Crosswalks are available (http://www.editeur.org/96/ONIX-and-MARC21)
  • OCLC, LOC, BDS all do this extensively. LOC is actively seeking more ONIX data from publishers

• greater use of ONIX in the discovery layer could give libraries and their patrons richer and much earlier visibility of the physical books and e-publications available to them
Jeff Penka

VP Product at Zepheira and The Library.Link Network
Becoming “Data Native”
A Path to Contemporary Relevance

Jeff Penka
Zepheira and The Library.Link Network
No...

Back to first principles...
Strengths can be source of challenges

Yeah, that seems like a great idea, but what about…
Becoming “Data Native”...

- Work with Intention
- Use Flexible, Modern Frameworks
- Understand primary, secondary, and edge use cases
What’s Your Motivation and Perspective?

Supplier (Push)  
Consumer (Pull)
Artificial Intelligence, Machine Learning, and the Web of Data

1. Revenue generated from the direct and indirect application of AI software will grow from $1.38 billion in 2016 to $59.75 billion by 2025.

2. The artificial intelligence (AI) market in the United States education sector is expected to grow at a compound annual growth rate of 47.5 percent during the period 2017-2021.

3. 2016 was a huge year for artificial intelligence.
   - Amazon Echo was a top-selling holiday gift, selling 5.2 million units (more than 2x the number sold in 2015).
   - IBM Watson diagnosed cancer.
   - Google’s DeepMind’s system, AlphaGo, cracked the ancient, complex Chinese game Go.
• Former chief scientist at Baidu,
• led the company's Artificial Intelligence Group
• Adjunct professor at Stanford University

Andrew Ng
@AndrewYNg

For AI to be free we need not just Open Source, but also a strong Open Data movement.

RETTWEETS 1,056
LIKES 1,657
1:06 PM - 5 Feb 2017

Rogue J. Bigham @jeffbigham · Mar 12
Replying to @AndrewYNg
perhaps. but, arguably, the data is much harder and expensive to collect than developing the algorithms to run on it

Michiel Trimpe @mtrimpe · Feb 5
Replying to @AndrewYNg
And strong data ownership. Google has every search click I ever made... but even though they’re mine I can’t get those anymore.
Purpose Driven, Audience Focused

- Need
- Manage
- Transform
- Find
- Measure
- Use
Becoming “Data Native”: Shared Vocabularies are Key Enablers

• In order to participate in broader conversations, we must speak so that others can act

• Data provides:
  • waypoints
  • connections
  • broader context
  • pathfinders

• We WILL NOT anticipate how it may be used
“Data Native” is about connections and transforming the way we & our data works

1. Source
   Data & Context

2. Transformation
   
   The Principle of Relativity
   - Instance
   - Book
   - (Item)
   - Print
   - Albert Einstein (author)
   - Theoretical Physics - Albert Einstein (concepts)

3. Reconciliation
   
   Einstein: His Life and Universe
   - Instance
   - Book
   - (Item)
   - Print
   - Walter Isaacson (author)
   - Albert Einstein (concept)

4. Network Connection
   
   Connect/Link Resources across the Network (Common & Unique)

5. Web Connections
   
   Connect/Link Resources across the Web (Understood Context)

MARC Records

Linked Data Resources

Consolidate “Shared” Linked Data Resources and Authorities

245: The Principle of Relativity
   100e: Albert Einstein
   6xx: Theoretical Physics
   Albert Einstein
   020: 978-0486600819

245: Einstein: His Life and Universe
   100e: Walter Isaacson
   6xx: Albert Einstein
   Albert Einstein
   020: 978-0743264747
Wikipedia = 20-50 Gig

Library.Link Network=
- 1500+ Library Systems
- 4000 Physical Locations
- 100,000+ Gig/
- 100+ TB Connected Data

>10%
1. Direct Use at Library:
   • Website
   • Catalog
   • Social

2. Direct Use in community:
   • Schools
   • Theater
   • Government
   • Personal

3. Direct & Organic:
   • Harvesters
   • Syndication Partners
With library resources accessible via published Linked Data, the Internet Archive can provide recognized reach to new audiences.
Thank You!