Discovery and Beyond: Leveraging Content and Big Data Analytics for Better Discovery

Eddie Neuwirth, Product Manager – Discovery and Delivery
Christine Stohn, Product Manager – Discovery and Delivery
AGENDA

- 2017 Roadmap Highlights & Content Update
  *What’s new?*

- Content Ingestion Process
  *How we get good Data*

- Enhancing Discovery with Intelligence from Big Data Analytics
  *What we do with the Data (now and future)*
Roadmap and Content Updates
2016-2017 Plan - Moving to a Shared Content and Data Model

- **Shared electronic collections**
- **Unified content ingestion process**
- **Loading of existing content and data enrichment to Primo Central / Summon Index**

No change to the method of managing local collections in Primo and Summon

- >90 Content types
- >4,000 Databases
- >2.6 B records
Immediate Benefits

Streamlined process for adding new content

Increased velocity for getting new and backlog content indexed

Provides opportunities to further enrich metadata and correct issues with metadata
2016-2017 – Progress Report

- Establish single ingestion pipeline for new content
- Consolidated Provider Agreements & Closed Content Gaps
- NERS voting & Content Idea Exchange

New Providers: 75
New Collections: 520
Commercial package updates & reloads: 1145
New packages: 73
Purged Packages: 176
2017 Content Roadmap

HIGH LEVEL PRIORITIES

Gap Analysis Project
- ~ 20 providers that have currency, metadata or linking issues
- Collection level audits

New Content
- NERS voting
- Idea Exchange
- Regional Content

Strategic Content
- Institutional Repositories
- Open Access
- Reference Content
- Ornid ID
- Controlled Vocabularies
Provider Optimization Program

• In partnership with our Global Content Alliance team established “P.O.P.”

• Quarterly meetings with our top content providers
  • Share our roadmaps and discuss new features that might impact content providers or that might require deeper metadata
  • Align on new content, new platforms, etc.
  • Perform content audits

• Mutual goal of optimizing and maximizing content and metadata available for discovery
Strategic and Continuous Content Growth

> 3 Billion items - Hundreds of Updates

New & Updated Content  Recent Highlights

Commercial Content Providers

- American Psychiatric Publishing
- Oxford University Press - Very Short Introductions and multiple collections
- CAIRN – Journals
- ProQuest - History Vault collections
- McGraw-Hill – Access Science and multiple collections
- SAGE - Knowledge Business Cases, Video and multiple collections
- PierOnline – eBooks (Japanese)
- Central and Eastern European Online Library (CEEOL) – Books and Journals
- Cambridge University Press - Ben Johnson Online and multiple collections
- MedicalOnline – (Japanese books)
- NBD Biblion Uittrekselbank (Dutch)
- Verlag C.H. Beck – (German books)
- McGraw-Hill - F.A. Davis PT Collection
- McGraw-Hill - HarrisonMedicina

- Gale – National Geographic, Business Insights, and Multiple Collections
- Henry Stewart Talks – Journals
- Edward Elgar Publishing Ltd - Reference Books
- Association for Computing Machinery (ACM)
- Elsevier - ClinicalKey ES guides and Techniques
- Chemical Society of Japan
- SOLIS
- SPIE Ebooks
- OECD (The Organisation for Economic Co-operation and Development)
- Data-Planet – Data sets
- ASTM Journals
- ProQuest Statistical Insight
- GESIS (German research repository)
Strategic and Continuous Content Growth

> 3 Billion items - Hundreds of Updates

New & Updated Content  Recent Highlights

Open Access & Open Educational Resources

- U.S. Department of Energy- OSTI SciTech Connect
- Parliamentary Library, Australia
- Exeley, Inc. – Open Access Journals
- OpenStax ebooks
- OpenSUNY Textbooks
- Open Textbook Library
- London Southbank Institute Open Research
- World Bank Open Knowledge Repository
- German National Library of Economics – EconStor
- Cognitive Sciences ePrint Archive
- 1Science – Oafinder
- ORBI - Université de Liège
- Public Library of Science (PLOS)

Open Access
Open Access Indicator and Filter – Beta Launch

>220 million Open Access items flagged
~75 providers
What do we mean by Open?

An Evolving & Ongoing Effort

GOALS:
• Make “openly available” content discoverable (not necessarily “Open Access”)

CHALLENGES:
• Metadata quality & accuracy
• Activation, linking, proxies, frequency of updates, and more

NEXT STEPS:
• Institutional Repositories
• Digital Collections

Learn More
Central Knowledge Center

Community Feedback Needed
Join our basecamp discussion
“Open Access special interest group”
Contact eddie.neuwirth@exlibrisgroup.com
Before We Begin...

1. Data: Garbage In; Garbage Out

“On two occasions I have been asked, ‘pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?’ ... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.”

Charles Babbage, “Passages from the Life of a Philosopher”

https://www.slideshare.net/CAinc/ten-commandments-for-tackling-fraud-the-role-of-bigdataandpredictiveanalytics
<table>
<thead>
<tr>
<th>Scope of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0 Billion Records in the Summon Index</td>
</tr>
<tr>
<td>1.4 Billion Records in the Primo Central Index</td>
</tr>
<tr>
<td>29.6 Million Titles in the 360 Knowledgebase</td>
</tr>
<tr>
<td>15.6 Million Portfolios in the Alma CZ</td>
</tr>
<tr>
<td>16,500+ Full-Text Databases in the 360 Knowledgebase</td>
</tr>
<tr>
<td>6,300+ Full-Text eCollections in the Alma CZ</td>
</tr>
<tr>
<td>Over 5,000 Data Sources, including Publishers, Vendors, and Content Providers</td>
</tr>
<tr>
<td>Managed and maintained by a team of 60+ Metadata Librarians, and Content Analysts worldwide</td>
</tr>
</tbody>
</table>
Content Ingestion Process for Discovery – High-Level Overview

1) Identify New Content to add to the Indexes
2) Review & Evaluate Sample Content
3) Approve Sample Content
4) Sign Data-Sharing Agreement
5) Acquire/Deliver Full Content Set
6) Data Standardization, Mapping, Link Rules, and Enrichment by the Metadata Librarians
7) Deduplication (Match & Merge)
8) Indexing
9) Post-Ingestion Maintenance and Support
# The Metadata That We Look For — Discovery

## Title Metadata
- Article Titles
- Book Titles
- Chapter Titles
- Publication Titles
- Subtitles

## Identifier Metadata
- Unique IDs for Specific Articles, Chapters, etc.
- Publication-level Unique Identifiers (e.g., ISSN, ISBN-10/13)
- Additional Identifiers (OCLC Number, LCCN, DOI, etc.)

## Publication Information Metadata
- Publisher
- Author(s)
- Corporate Author(s)
- Volume Numbers
- Issue Numbers
- Start Pages
- Publication Date(s)
- Publication Series or Imprints
- Publication Place

## Additional Metadata
- Subject Headings
- Keywords
- Language
- Classification
- Abstracts
- Open Access Flags
- Full-Text Available Flags

## Metadata Schema & Structure
- MARC21
- MODS
- Dublin Core
- NLM / JATS
- ONIX
- Well-structured, and labeled, homegrown schemas
Example of a Good Metadata Record (Discovery)

• The metadata highlighted in red are the metadata points we typically look for when reviewing sample data.

• The metadata here is fairly complete, and includes several unique identifiers such as both print and electronic ISBNs, and an OCLC control number.

• The metadata fields are well-structured, labeled, and in a logical order that is easy to read and understand.

```
<eBook>
  <Content_Modified>false</Content_Modified>
  <BibData>
    <EISBN>9781614581048</EISBN>
    <PISBN>9780892216963</PISBN>
    <OCLC_Control>817084723</OCLC_Control>
    <Title>Yesterday, Today and Forever</Title>
    <Publication_Date>2009-06-01 00:00:00.0</Publication_Date>
    <Imprint>New Leaf Publishing Group, Inc.</Imprint>
    <Pages>177</Pages>
    <Adult_Juvenile>A</Adult_Juvenile>
    <Product_ID>10626803</Product_ID>
  </BibData>
  <Contributor>
    <Name>von Trapp, Maria</Name>
    <Role>Author</Role>
  </Contributor>
  <BISAC>
    <Literal>RELIGION</Literal>
  </BISAC>
  <BISAC>
    <Literal>Christian Life / Inspirational</Literal>
  </BISAC>
</eBook>
```
Example of a More Challenging Metadata Record (Discovery)

<source-issue>
  <key>0002457632</key>
  <T1>N 20161028</T1>
  <PW>IP200722</PW>
  <GA>158AN</GA>
  <SQ>13382J0</SQ>
  <PT>J</PT>
  <SO>ACTA PETROLOGICA SINICA</SO>
  <JI>Acta Petrol. Sin.</JI>
  <J1>ACTA PETR S</J1>
  <J2>ACTA PETROL SIN</J2>
  <J9>ACTA PETR SIN</J9>
  <CF>D</CF>
  <SC>D KY GEOLOGY</SC>
  <SN>1000-0569</SN>
  <PU>SCIENCE PRESS</PU>
  <PI>BEIJING</PI>
  <PA>16 DONGHUANGCHENGGEN NORTH ST, BEIJING 100717, PEOPLES R CHINA</PA>
  <VL>23</VL>
  <IL>00002</IL>
  <LD>20070528-14:53:44</LD>
  <IO>20160002457632</IO>
</source-issue>

<source-item>
  <key>000245763200001</key>
  <T2>C 20161023</T2>
  <T3.AR.Cr/T3>
  <TI>Lu-Hf isotopic systematics and their applications in petrology</TI>
  <AU>Wu, FY</AU>
  <RO>Author, Reprint Author</RO>
  <LN>Wu F Y</LN>
  <EM>wufuyuan@mail.igcas.ac.cn</EM>
  <AU>Li, XH</AU>
  <RO>Author</RO>
  <LN>Li X H</LN>
  <AU>Zheng, YF</AU>
  <RO>Author</RO>
  <LN>Zheng Y F</LN>
  <AU>Gao, S</AU>
  <RO>Author</RO>
  <LN>Gao S</LN>
  <DT>R Review</DT>
  <BP>185</BP>
  <EP>220</EP>
  <PG>36</PG>
  <LA>CH Chinese</LA>
  <DE>zircon</DE>
  <DE>whole-rock</DE>
  <DE>Lu-Hf isotope</DE>
  <DE>petrology</DE>
  <ID>ZIRCON U-PB</ID>
  <ID>NORTH CHINA CRATON</ID>
  <ID>MC-ICP-MS</ID>
  <ID>PLASMA-MASS SPECTROMETRY</ID>
  <ID>N-SITU HF</ID>
  <ID>MANTLE HETEROGENEITY BENEATH</ID>
  <ID>ULTRAHIGH-PRESSURE ECLOGITE</ID>
  <ID>OCEAN FERROMANGANESE CRUSTS</ID>
  <ID>TRACE-ELEMENT FRACTIONATION</ID>
  <ID>SPINEL PERIDOTITE XENOLITHS</ID>
  <AV>Y</AV>
  <NR>303</NR>
</source-item>
Determining Linking Methods

Index-Enhanced Direct Linking (IEDL)
- Summon constructs direct link rules to the publisher/provider platforms based on the metadata supplied
- IEDL tends to be more stable and successful when getting users to the content that they are looking for

OpenURL Linking
- Summon and Primo Central send the metadata to a link resolver (360 Link, SFX, etc.)
- The link resolver will then attempt to create a URL that resolves to the target platform
## Data Standardization, Mapping, Link Rules, and Enrichment - Discovery

### Data Standardization
- Very basic high-level clean-up of the data to standardize it
- Examples include:
  - Remove leading/trailing white spaces in Title and Subtitle fields
  - Clean-up diacritics and other encoding issues

### Mapping
- Metadata Librarians map the metadata fields in the records to our schemas
- This allows the metadata to appear in the UI and/or be made searchable within Summon or Primo Central

### Link Rules
- Metadata Librarians will create or modify link rules to enable direct linking of the content from Summon

### Enrichment
- Enriching the content by adding additional metadata when applicable
- Examples:
  - Scholarly/peer-reviewed flags from Ulrichsweb ( Summon)
  - Citation counts from Scopus ( Summon)
  - Citation trails ( Primo Central)
Match & Merge at a High-Level

- Summon displays a Single Record
- Primo displays an enriched record and options to view all versions

Cancer: The RAS Renaissance
by Ledford, Heidi

*Nature*, ISSN 0028-0836, 04/2015,
Volume 520, Issue 7547, pp. 278-280
Match & Merge Rules

ISSN Match Rule

- ISSN
- Title
- Publication Date
- Content Type

= Match & Merge

ISBN Match Rule

- ISBN
- Title
- Publication Date

= Match & Merge

DOI Match Rule

- DOI
- Fuzzy Title

= Match & Merge
Match & Merge Challenges

• Slight differences in the metadata can cause records to not de-duplicate

• Each provider/source will treat the metadata differently, making it difficult at times to reconcile or catch upfront

• Make the rules too strict, and you’ll undermatch; make the rules too loose, and you’ll overmatch
Post-Ingestion Maintenance & Support

Currency

- Currency is the process of the publisher/provider sending to Summon new/updated metadata records, or record deletions for content that need to be removed

- Frequency of providing updates is often at the discretion of the publisher/provider

Metadata Issues

- Address reported issues of metadata quality from Summon customers

- Most issues involve incorrect metadata, or slight variations in the metadata that may impact OpenURL linking or the Match & Merge process and create duplicate records
Big data analytics – features for a richer discovery experience
Enhancing discovery with intelligence from big data analytics

Big Data Analysis

- Databases & Full Text
- Metadata & Taxonomies
- Search Logs
- Enhanced searching
- New Research Paths
- Richer Learning Experience
Artificial intelligence and big data analytics in discovery

• Enriching searches and enriching data

• Understanding data context and building relationship (knowledge) graphs

• Provide new discovery services beyond searching
Search expansion with controlled vocabulary

Search expansion with controlled vocabulary

177,447 results sorted by relevance
including "attention deficit hyperactivity disorder"

Including "asata shakur". Just search asata shakur

PAGE 1 257 Results  Personalize
Big data analytics – features for a richer discovery experience

Search expansions

Citation trail

Children's social behaviour, language and

Children's Language and Behavioural, Social and

Child and informant influences on behavioral

NOW LISTING SOURCES CITED IN:
Teacher and observer ratings of head start children's social skills
Millport, Rollie; Greenfield, Daryl B
Browsing through citation lists

Your current citation path

1. Investing in the transition to sustainable agriculture
2. Transforming U.S. Agriculture
3. Agroecology and Alternative Agri-food Movements in ...
4. Agroecology and Sustainable Livelihoods ...
5. NOW LISTING SOURCES CITING:

Agroecology versus input substitution: A fundamental contradiction of sustainable ... Rosset, Peter M.; Altieri, Miguel A.

PAGE 1  26 Results

1. ARTICLE / multiple sources exist, see all
   What Is Sustainable Agriculture? A Systematic Review
   Velten, Sarah; Leventon, Julia; Jager, Nicolas; Newig, Jens 2015
   Sustainability, Vol.7(6), pp.7833-7865 [Peer Reviewed Journal]
   politics and policies of sustainable agriculture: Some... aspects associated with sustainable agriculture: (2)
   detect... sustainable agriculture is very ... Full text available

2. ARTICLE / multiple sources exist, see all
   Agroecology and Sustainable Livelihoods: Towards an Integrated Approach to Rural Development
   Ameikawa, Yuichiro
   Agroecology; Sustainable Livelihoods; Sustainable Agriculture; Integrated Approach To Rural Development; Small-Scale Farmers
   ... and G. Bedi (eds.), 2002. Sustainable agriculture and... sustainable agriculture and rural development... Journal of Sustainable Agriculture,
   Full-text not available. Click on the "View It" tab for additional options
Big data analytics – features for a richer discovery experience

Search expansions

Citation trail

bX Recommender
Development of a framework for knowledge translation: understanding user context

Jacobson, Nora; Butterill, Dale; Goering, Paula

Health Services Research; Information Dissemination; Knowledge: Information Services -- Utilization

Journal of health services research & policy, April 2003, Vol.8(2), pp.94-9[Peer Reviewed Journal]

Full text available

Related reading

Demystifying knowledge translation: learning from... Patricia Martens

A guide to knowledge translation theory. Carole A Estabrooks

How can research organizations more... John N Lavis
The Milbank quarterly., Vol. 81(2), p. 221-48
Tragedy and ethical agency in *Hegel's The Spirit of Christianity* and Its Fate
by Thibodeau, Martin
Philosophy & Theology, 09/2012, Volume 24, Issue 2

It performs this task on the basis of a reading of Hegel's early essay *The Spirit of Christianity* and Its Fate.

Journal Article: Full Text Online
Preview ▼ Related Articles ▼

Forgiveness, freedom, and human finitude in *Hegel's The Spirit of Christianity* and Its Fate
by Theodore, George
International Philosophical Quarterly, 03/2011, Volume 51, Issue 1

The purpose of this essay is to consider the significance that Hegel grants to religious love and, with it, forgiveness in his early *The Spirit of Christianity* and Its Fate.

Journal Article: Full Text Online
Preview ▼ Related Articles ▼

Related Articles

- Tragedy and Understanding in Hegel's Dialectic
  by Lumsden, Simon

- "The Spirit of Christianity and Its Fate" Towards a... (partial)
  by Ormiston, Alice
  Canadian journal of political science. 2002 ,Vol. 35 (3) p. 499-525

- Hegel and the Search for Epistemological Criteria
  by Ward, Andrew
  Idealistic studies. 1992 ,Vol. 22 (3)

- Love is colder than death: negativity, infinity and... (partial)
  by Saffati, Y.
Big data analytics – In the research lab

Controlled vocabulary
- Purpose
  - Search expansions
  - Topic exploration features
  - “Did you mean” enhancements
  - Search and ranking improvements
  - Creating search to discipline match

Document relations
- Books and reviews
- Articles and research data
- ...

Topic relations
- Topic exploration paths
<table>
<thead>
<tr>
<th>#</th>
<th>Title</th>
<th>Authors</th>
<th>Location</th>
<th>Date</th>
<th>Pages</th>
<th>Status</th>
<th>Full text available</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Search for Authority in Reformation Europe</td>
<td>Fulton, Elaine; Parish, Helen; Webster, Peter, Dr.; Cameron, Evan, Professor; Gordon, Bruce, Professor; Heal, Bridget, Dr.; Mason, Roger A, Professor; Nelson Burnett, Amy, Professor; Pettigrew, Andrew, Dr.; Von Greyerz, Kaspar, Professor; Ashgate 2014</td>
<td>Authority -- Religious Aspects -- Christianity -- History Of Doctrines ; Protestant Churches -- Doctrines: Reformation -- Europe</td>
<td>2014</td>
<td></td>
<td>Books</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>St Andrews Studies in Reformation History: Search for Authority in Reformation Europe</td>
<td>Fulton, Elaine; Dr; Parish, Helen; Fulton, Elaine; Webster, Peter</td>
<td>Farnham, GB: Ashgate Publishing Ltd 28 February 2014 History ; Medieval Authority in Reformation Europe a matter of utmost... Search for Authority in Reformation Europe 89 The... The Search for Authority in Reformation Europe...</td>
<td>2014</td>
<td></td>
<td>Books</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>Oxford Historical Monographs: Pius IV and the Fall of the Carafa: Nepotism and Papal Authority in Counter-Reformation Rome</td>
<td>Pattenden, Miles</td>
<td>Oxford, GBR: OUP Oxford 01 February 2013 History ; Europe – General Udo-Renate Blumenthal and Guy Fitch Lytle (eds), Reform and Authority in the Medieval and Reformation</td>
<td>2013</td>
<td></td>
<td>Books</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Search for Authority in Reformation Europe

Fulton, Elaine; Parish, Helen; Webster, Peter, Dr.; Cameron, Euan, Professor; Gordon, Bruce, Professor; Heal, Bridget, Dr.; Mason, Roger A, Professor; Nelson Burnett, Amy, Professor; Petegree, Andrew, Dr.; Von Greyerz, Kaspar, Professor.

Ashgate 2014 Authority -- Religious Aspects -- Christianity -- History Of Doctrines; Protestant Churches -- Doctrines; Reformation -- Europe;

Full text available

3 Book reviews available
<table>
<thead>
<tr>
<th>#</th>
<th>Title</th>
<th>Authors</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Search for Authority in Reformation Europe</td>
<td>Fulton, Elaine; Parish, Helen; Webster, Peter, Dr; Cameron, Euan, Professor; Gordon, Bruce, Professor; Heal, Bridget, Dr; Mason, Roger A, Professor; Nelson Burnett, Amy, Professor; Pettegree, Andrew, Dr; Von Greyerz, Kaspar, Professor;</td>
<td>Book reviews for:</td>
</tr>
<tr>
<td>5</td>
<td>The Search for Authority in Reformation Europe (Book Review)</td>
<td>Rittgers, Ronald K.</td>
<td>Visual Arts; History &amp; Archaeology; Drama; Languages &amp; Literatures; Renaissance Quarterly, 01 March 2015, Vol.68(1), pp.341-343[Peer Reviewed Journal]</td>
</tr>
<tr>
<td>6</td>
<td>Helen Parish, Elaine Fulton and Peter Webster, eds, The Search for Authority in Reformation Europe</td>
<td>Atherton, Ruth</td>
<td>Parish, Helen; Fulton, Elaine; Webster, Peter; European History Quarterly, 2015, Vol.45(3), pp.589-591[Peer Reviewed Journal]</td>
</tr>
</tbody>
</table>

Full text available

"turn was to trigger the World War. Helen Parish, Elaine Fulton and Peter Webster, eds, The Search for Authority in Reformation Europe"
Big data analytics – In the research lab

Controlled vocabulary
- Purpose
  - Search expansions
  - Topic exploration features
  - “Did you mean” enhancements
  - Search and ranking improvements
  - Creating search to discipline match

Document relations
- Books and reviews
- Articles and research data
- ...

Topic relations
- Topic exploration paths