Content Ingestion: The Trials and Triumph

Primo & Summon Content Updates + Big Data Analytics for Better Discovery

Eddie Neuwirth, Product Manager – Discovery and Delivery
Dave Hovenden, Sr. Manager – Content Operations
Christine Stohn, Product Manager – Discovery and Delivery
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: 520
New Collections: 75
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
• Orcid 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
Content Ingestion: The Trials and Triumph

Dave Hovenden – Sr. Manager, Content Operations
Content Ingestion: The Trials and Triumph

ELUNA 2017 – May 11, 2017

Dave Hovenden – Sr. Manager, Content Operations
Before We Begin...

- 2.9 Billion Records in the Summon Index
- 1.3 Billion Records in the Primo Central Index
- 29.6 Million Titles in the 360 Knowledgebase
- 15.6 Million Portfolios in the Alma CZ
- 16,500+ Full-Text Databases in the 360 Knowledgebase
- 6,300+ Full-Text eCollections in the Alma CZ
- Over 5,000 Data Sources, including Publishers, Vendors, and Content Providers
- Managed and maintained by a team of 60+ Metadata Librarians, and Content Analysts worldwide
Discovery Content Ingestion
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

<table>
<thead>
<tr>
<th>Title Metadata</th>
<th>Identifier Metadata</th>
<th>Publication Information Metadata</th>
<th>Additional Metadata</th>
<th>Metadata Schema &amp; Structure</th>
</tr>
</thead>
</table>
| • Article Titles  
• Book Titles  
• Chapter Titles  
• Publication Titles  
• Subtitles | • Unique IDs for Specific Articles, Chapters, etc.  
• Publication-level Unique Identifiers (e.g., ISSN, ISBN-10/13)  
• Additional Identifiers (OCLC Number, LCCN, DOI, etc.) | • Publisher  
• Author(s)  
• Corporate Author(s)  
• Volume Numbers  
• Issue Numbers  
• Start Pages  
• Publication Date(s)  
• Publication Series or Imprints  
• Publication Place | • Subject Headings  
• Keywords  
• Language  
• Classification  
• Abstracts  
• Open Access Flags  
• Full-Text Available Flags | • 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.
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>
  <J2>ACTION PETROL SIN</J2>
  <J9>ACTION PETR SIN</J9>
  <CF>D</CF>
  <SC>D KY GEOLOGY</SC>
  <SN>1000-0569</SN>
  <PU>SCIENCE PRESS</PU>
  <PL>BEIJING</PL>
  <PA>16 DONGHUANGCHENGGEN NORTH ST, BEIJING 100717, PEOPLES R CHINA</PA>
  <VL>23</VL>
  <IS>2</IS>
  <PY>2007</PY>
  <PD>FEB</PD>
  <TV>Y</TV>
  <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>
  <T1>Lu-Hf isotopic systematics and their applications in petrology</T1>
  <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>MICP-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 PERIODITE 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)
• Summon displays a Single Record
• Primo displays an enriched record and options to view all versions
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
Knowledgebase Content Ingestion
Content Ingestion Process for KB – High-Level Overview

1) Identify New Content to add to the Knowledgebases
2) Review & Evaluate Sample Content
3) Approve Sample Content
4) Acquire/Deliver Full Content Set
5) Create the Database in the Knowledgebase
6) Data Standardization, Mapping, and Enrichment by the Metadata Librarians
7) Load Title List into the Database within the Knowledgebase
8) Set-up the Linkers
9) Post-Ingestion Maintenance and Support
**The Metadata That We Look For — Knowledgebase**

<table>
<thead>
<tr>
<th>General</th>
<th>Serials</th>
<th>Books</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Title</td>
<td>• Start Date</td>
<td>• Date Published (Print)</td>
</tr>
<tr>
<td>• Identifier (Print)</td>
<td>• End Date</td>
<td>• Date Published (Online)</td>
</tr>
<tr>
<td>• Identifier (Online/Electronic)</td>
<td>• Embargo</td>
<td>• Author</td>
</tr>
<tr>
<td>• URL</td>
<td>• Volume*</td>
<td>• Volume</td>
</tr>
<tr>
<td>• Coverage (Full-Text vs. Abstracts)</td>
<td>• Issue*</td>
<td>• Edition</td>
</tr>
<tr>
<td>• Content Type</td>
<td></td>
<td>• Editor</td>
</tr>
<tr>
<td>• Title ID</td>
<td></td>
<td>• Parent Publication Title ID</td>
</tr>
<tr>
<td>• Publisher</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Access Type (Paid or Free)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• We **highly recommend (and prefer)** that publishers and providers supply their data in the Knowledge Bases And Related Tools (KBART) format

*SFX and Alma CZ KBs only*
Data Standardization, Mapping, and Enrichment - Knowledgebase

Data Standardization

- Basic clean-up of the data before being ingested into the Knowledgebase
- Examples include:
  - Clean-up diacritics and other encoding issues
  - Date formatting

Data Mapping

- Metadata Librarians map the metadata fields in the records to the Knowledgebase schemas

Data Enrichment

- Enriching the content by adding additional metadata when applicable
- Examples:
  - Represent Title Splits
  - Original/Copy MARC Cataloging
  - Authority Record creation
The 360 KB team has its own in-house team of Metadata Librarian Catalogers that maintain and normalize provider holdings to 360 KB Authority Records.

Holdings for the same title, regardless of who the provider is, are all mapped and linked to an Authority Record (when available) and assigned a Serials Solutions ID (SSID).

Authority Records may be based off MARC records from the Library of Congress, National Library of Medicine, Bowker, or the provider itself.

The Catalogers also do original/copy cataloging of MARC records for the Library of Congress and the Program for Cooperative Cataloging (PCC).

The 360 KB Catalogers are the only commercial vendor partner participating in the Electronic Cataloging in Publication (ECIP) Program.
Post-Ingestion Maintenance & Support

Metadata Issues
- Add ISBN or ISSN
- Dates of Coverage
- Incorrect/Bad Metadata
- Missing Title(s)
- Extra Title(s)
- Broken URLs (Journal and/or Database-level)
- Title Split Request
- Normalization Request

Provider/Database Issues
- Platform Migrations
- Database Retirement
- Improperly Defined Provider Collections

Currency
- Perform regular/routine database updates
- Automated holdings updates from select providers
How the 360 Knowledgebase and Summon Work Together for Discovery

• Summon relies upon the 360 KB to help facilitate rights access to the content that resides within the index

• Rights access is assigned by tracking a particular title by its ISSN or ISBN in 360 KB, or by assigning a Database ID (all or nothing rights)

• Summon utilizes the SSID found within the Authority Records by adding it to any records found within the Summon Index that match based on either the ISSN or ISBN

• The 360 KB also helps Summon indicate when content has full-text availability
Some of the Challenges

- Metadata Quality
- Lack of Consistent Metadata Standard for Discovery
- Reconciling Differences between Providers for the Same Piece of Content
- Provider Co-operation and/or Responsiveness
- Understanding a Provider’s Subscription Model for Rights and Linking
- Acquisition/Delivery Issues
- Volume of Content/Data
- Number of Providers/Sources
To Reiterate...

2.9 Billion Records in the Summon Index
16,500+ Full-Text Databases in the 360 Knowledgebase

1.3 Billion Records in the Primo Central Index
6,300+ Full-Text eCollections in the Alma CZ

29.6 Million Titles in the 360 Knowledgebase
Over 5,000 Data Sources, including Publishers, Vendors, and Content Providers

15.6 Million Portfolios in the Alma CZ
Managed and maintained by a team of 60+ Metadata Librarians, and Content Analysts worldwide
(A Triumph in Progress)

Another brief word from Eddie

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
Big data analytics – features for a richer discovery experience

Christine Stohn–
Product Manager
Enhancing discovery with intelligence from big data analytics

- Databases & Full Text
- Metadata & Taxonomies
- Search Logs

Big Data Analysis

- Enhanced searching
- New Research Paths
- Richer Learning Experience
Big data analytics – features for a richer discovery experience

Search expansions

177,447 results sorted by relevance

including "attention deficit hyperactivity disorder"

Just search adhd
Big data analytics – features for a richer discovery experience

Search expansions

Citation trail

© 2017 Ex Libris | Confidential & Proprietary
Browsing through citation lists

Investing in the transition to sustainable agriculture
Transforming U.S. Agriculture
Agroecology and Alternative Agrofood Movements
Agroecology and Sustainable Livelihoods

NOW LISTING SOURCES CITING:

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

What Is Sustainable Agriculture? A Systematic Review
Veltman, 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

Agroecology and Sustainable Livelihoods: Towards an Integrated Approach to Rural Development
Ameikawa, Yulichiro
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
36,935 results sorted by relevance

1. 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

2. Forgiveness, freedom, and human finitude in Hegel's *The Spirit of Christianity* and Its Fate
   by George, Theodore
   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

Related Articles

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

- "The Spirit of Christianity and Its Fate". Towards a...
  by Ormiston, Alice

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

- Love is colder than death: negativity, infinity and...
  by Saffle, V.
Big data analytics – In the research lab

**Controlled vocabulary**
- Baseline: LoC, Wikipedia, LC names
- 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
reformation europe authority

1. **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; Pettigrew, Andrew, Dr; Von Greyerz, Kaspar, Professor; Ashgate 2014
   Authority -- Religious Aspects -- Christianity -- History Of Doctrines ; Protestant Churches -- Doctrines; Reformation -- Europe;
   Full text available

2. **St Andrews Studies in Reformation History : Search for Authority in Reformation Europe**
   Fulton, Elaine, Dr; Parish, Helen; Fulton, Elaine; Webster, Peter
   Farnham, GB: Ashgate Publishing Ltd 26 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... Full text available

3. **Oxford Historical Monographs : Pius IV and the Fall of The Carafa : Nepotism and Papal Authority in Counter-Reformation Rome**
   Pattenden, Miles
   Oxford, GBR: OUP Oxford 01 February 2013
   History ; Europe -- General
   Uto-Renate Blumenthal and Guy Fitch Lytle (eds), Reform and Authority in the Medieval and Reformation
   Full text available
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; Pettigree, Andrew, Dr; Von Geyerz, Kaspar, Professor;

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

Reformation -- Europe;

Full text available

3 Book reviews available
Book reviews for:

1. **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; Pettegree, Andrew, Dr; Von Greyerz, Kaspar, Professor;

5. **The Search for Authority in Reformation Europe (Book Review)**
   Rittgers, Ronald K.
   Visual Arts; History & Archaeology; Drama; Languages & Literatures;
   Full text available

6. **Helen Parish, Elaine Fulton and Peter Webster, eds, The Search for Authority in Reformation Europe**
   Atherton, Ruth
   Parish, Helen; Fulton, Elaine; Webster, Peter;
   The decision 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
- Baseline: LoC, Wikipedia, LC names
- 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