

May 11, 2017

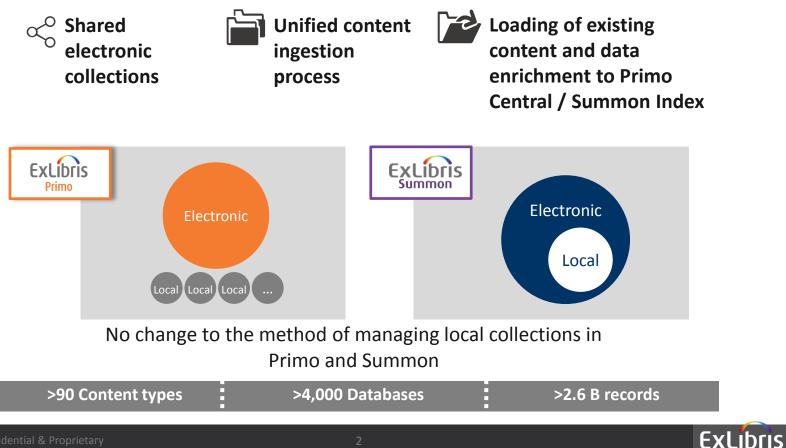
Content Ingestion: The Trials and Triumph

Primo & Summon Content UpdatesBig 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



Immediate Benefits

66 What We Promised 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

ExLibris Primo

> 75 New Providers 520 New Collections



\mathbf{F}

- **1145** Commercial package updates & reloads
- 73 New packages
- **176** Purged Packages



2017 Content Roadmap

HIGH LEVEL PRIORITIES







Gap Analysis Project

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

New Content

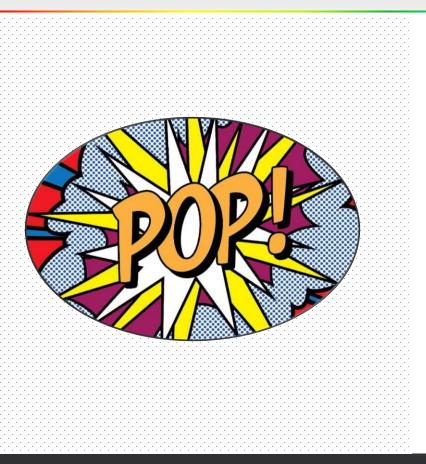
- NERS voting
- Idea Exchange
- Regional Content

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



Users of North Ame

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



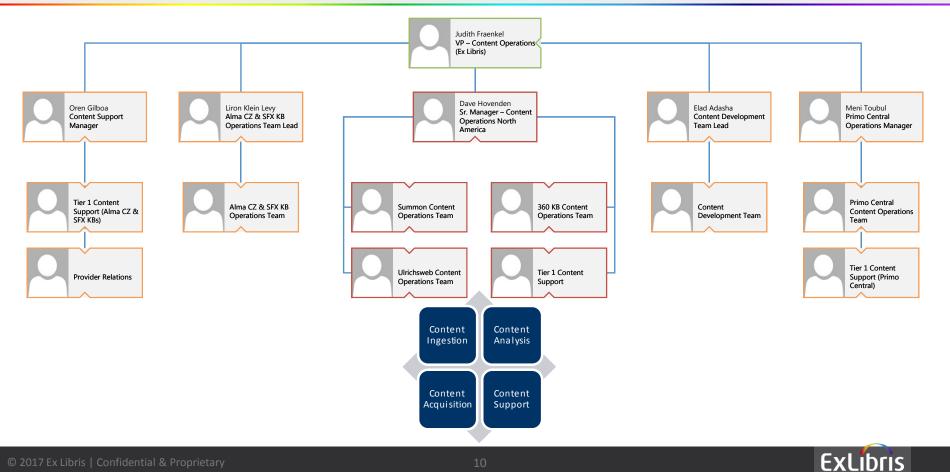
Over 5,000 Data Sources, including Publishers, Vendors, and Content Providers



Managed and maintained by a team of 60+ Metadata Librarians, and Content Analysts worldwide



Ex Libris Content Operations

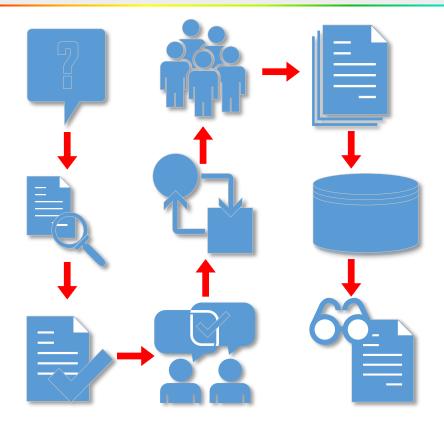


Discovery Content Ingestion

EXLIDIS. a ProQuest Company



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	Identifier Metadata	Publication Information Metadata	Additional Metadata	Metadata Schema & Structure
 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)

<eBook>

<Content_Modified>false</Content_Modified> <BibData> <**EISBN**>9781614581048</EISBN> <**PISBN**>9780892216963</PISBN> <**OCLC_Control**>817084723</OCLC_Control> <**Title**>Yesterday, Today and Forever</Title> <Sub_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>

- 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 wellstructured, 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> <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> <IS>2</IS> <PY>2007</PY> <PD>FEB</PD> $\langle TV \rangle Y \langle TV \rangle$ <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.CI</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> wufuyuan@mail.igcas.ac.cn <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>IN-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> $\langle AV \rangle Y \langle AV \rangle$ <NR>303</NR> <RS>DOI 10.1016/i.epsl.2005.07.011</RS> </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

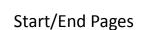
•Metadata Librarians will create or modify link rules to enable direct linking of the content from Summon

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

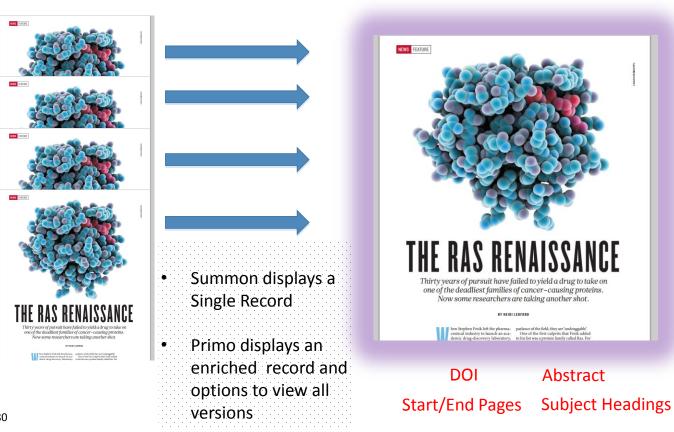


Subject Headings

Abstract

DOI

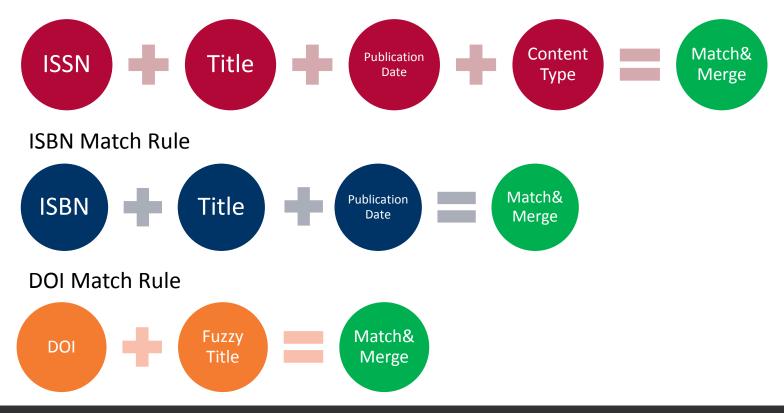




ExLibris

Match & Merge Rules

ISSN Match Rule



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

		Preview -			U	
3	online	Protective Effect of Tetracycline against Dermal Toxicity Induced by Jellyfish Venom by Kang, C: Jin, YB: Kwak, J: More	99	M	••••	
		PLOS ONE, 03/2013, Volume 8, Issue 3				
	0 Open Access	Background: Previously, we have reported that most, if not all, of the Scyphozoan jellyfish venoms contain multiple components of metalloproteinases, which apparently linked to the venom toxicity				
	opennous					
		Journal Article: Full Text Online	WEB OF SCI	ENCE		
		Preview -	5		U	
4	online	Protective Effect of Tetracycline against Dermal Toxicity Induced by Jellyfish Venom: e57658 by <u>Changkeun Kang; Yeung Bae Jin; Jeongsoo Kwak; More</u>	99	M	∎₀	
		PLoS One, 03/2013, Volume 8, Issue 3				
		Background Previously, we have reported that most, if not all, of the Scyphozoan jellyfish venoms contain multiple components of metalloproteinases, which apparently linked to the venom toxicity				
		Journal Article: Full Text Online				
		Preview Related Articles			0	
5	online	Protective Effect of Tetracycline against Dermal Toxicity Induced by Jellyfish Venom by Kang, Changkeum; Yeung Bae Jin; Kwak, Jeongsoo; More	99	M	i	
		PloS One, 2013, V. anne e, issue 3				
	Open Access	BackgroundPreviously, we have reported that most, if not all, of the Scyphozoan jellyfish venoms contain multiple components of metalloproteinases, which apparently linked to the venom toxicity				
		Journal Article: Full Text Online				
		Preview Related Articles			U	



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

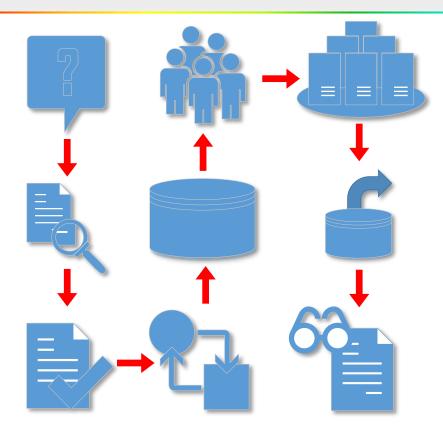
- Address reported issues of metadata quality from Summon customers
- Most issues involve incorrect metadata, or slight variations in the metadata that may impact OpenURL linking or the Match & Merge process and create duplicate records

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

General

- Title
- Identifier (Print)
- Identifier (Online/Electronic)
- URL
- Coverage (Full-Text vs. Abstracts)
- Content Type
- Title ID
- Publisher
- Access Type (Paid or Free)

*SFX and Alma CZ KBs only

Serials

- Start Date
- End Date
- Embargo
- Volume*
- Issue*

Books

- Date Published (Print)
- Date Published (Online)
- Author
- Volume
- Edition
- Editor
- Parent Publication
 Title ID

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





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



Knowledgebase Cataloging and Authority Records

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 <u>only</u> commercial vendor partner participating in the Electronic Cataloging in Publication (ECIP) Program



Program for Cooperative Cataloging

BIBCO / CONSER / NACO / SACO





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

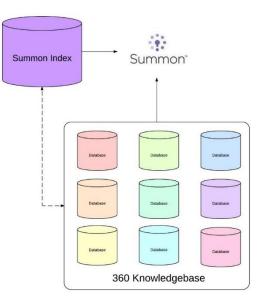
- Perform regular/routine database updates
- Automated holdings updates from select providers



ExLibris

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



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





(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

				Summon*	Ľ	gravitation waves	
University Library Back to top	Observation of Gra	witational Waves from a Binary Black Hole Merger Gravitational waves from resolvable massive black hole binary systems and observations with Pulsar Timing Arrays by Sesana, A: Vecchio, A: Volonteri, M Monthly Notices of the Royal Astronomical Society, 2009, Volume 394, Issue 4 and to gain insight into the relevant physical processes. Pulsar	0 99	REFINE YOUR SEARCH Full Text Online Scholarly & Peer-Review Peer-Review Open Access	/	Datab	ts sorted by <u>relevance</u> ー ふ base Recommendations Frust (Public Domain within US only) org Gravitation-Wave Emission in Shift-
	Open Access	Timing Arrays (PTAs) currently provide the only means to observe gravitational radiation from massive black hole binary systems with masses ≿107 M Journal Article: <u>Full Text Online</u> Preview ▼		Library Catalog CONTENT TYPE Book / eBook (81,761) Journal Article (36,393)			Theories by <u>Barausse, E; Yaqi, K</u> PHYSICAL REVIEW LETTERS, 11/2015, Volu Gravity theories beyond general relativity f gravitational emission by compact-star bin sourced by "sensitivity"
	12 online	Synthetic model of the gravitational wave background from evolving binary compact objects by <u>Dvorkin, Irina; Uzan, Jean-Philippe; Vangioni, Elisabeth; More</u> Physical Review D, 11/2016, Volume 94, Issue 10 Modeling the stochastic gravitational wave background from	22 i	Dissertation/Thesis (2,227) Book Chapter (1,482)			Journal Article: <u>Full Text Online</u> Preview •
		various astrophysical sources is a key objective in view of upcoming					ExLibris

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 http://bit.ly/Summon-OpenAccess

Community Feedback Needed Join our basecamp discussion "Open Access special interest group" Contact eddie.neuwirth@exlibrisgroup.com



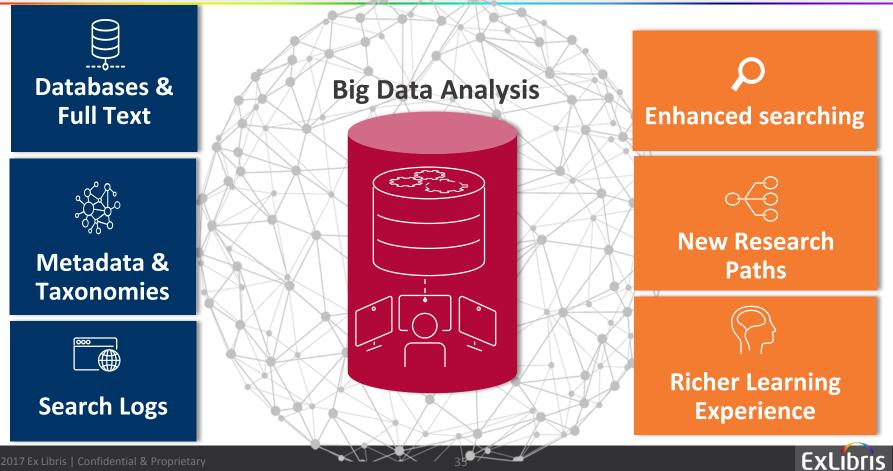


Big data analytics – features for a richer discovery experience

Christine Stohn– Product Manager



Enhancing discovery with intelligence from big data analytics

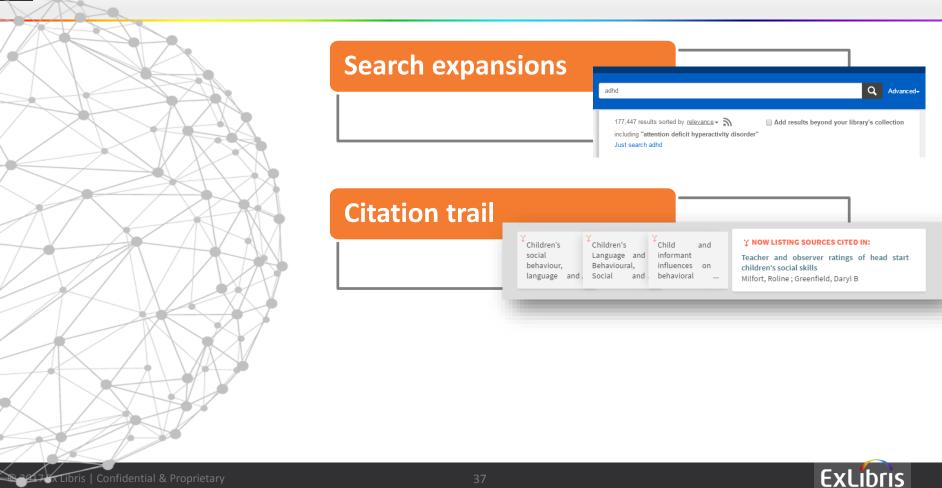


Big data analytics – features for a richer discovery experience

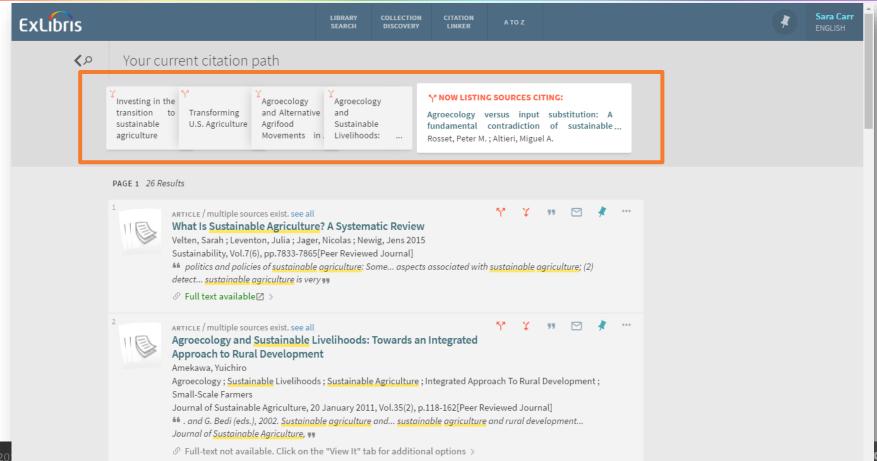
adhd	Q Advanced-
177,447 results sorted by <u>relevance</u> - S including "attention deficit hyperactivity disorder" Just search adhd	Add results beyond your library's collection



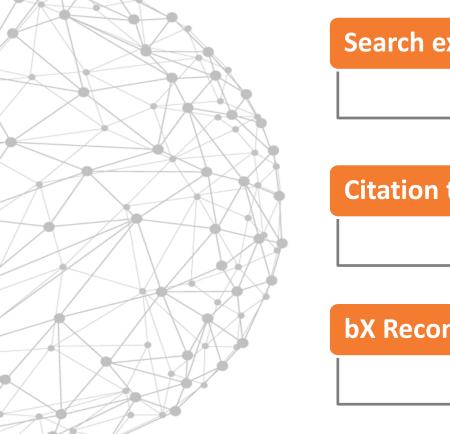
Big data analytics – features for a richer discovery experience



Browsing trough citation lists



Big data analytics – features for a richer discovery experience



Search expansions	
	adhd Q Advanc
	177,447 results sorted by relevance - A G Add results beyond your library's collection including "attention deficit hyperactivity disorder" Just search adhd
Citation trail	^Y Children's ^Y Child and Language and informant Teacher and observer ratings of head start
behaviour, language and .	Language and informant Teacher and observer ratings of head start Behavioural, influences on children's social skills Social and behavioral Milfort, Roline ; Greenfield, Daryl B
oX Recommender	bX Users interested in this article also expressed an interest in the following: About bX 1. Ridings, Catherine M <u>Virtual community attraction:</u> Why people hang out online Journal of S S P
	computer-mediated communication 10.1 (2004) No Pagination Specified 2. Barrett, H. Clark <u>Can Manipulations of Cognitive Load Be Used to Test Evolutionary</u>
	Hypotheses?_Journal of personality and social psychology 91.3 (2006) 513-518 3. COUZIN, JENNIFER <u>Friendship as a Health Factor</u> _Science 323.5913 (2009) 454-457
	4. Wood, Stephen The four pillars of HRM: are they connected? Human resource S & *
	5. Tarrant, Mark Friendship group identification, multidimensional self-concept, and experience of developmental tasks in adolescence. Journal of Adolescence 29.4 (2006) 627-640

ExLibris

bX Recommender on Primo



TOP

ARTICLE / multiple sources exist. see all 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]

𝔄 Full text available

SEND TO Send to VIEW IT B Đ ቡ ß ${\mathscr O}$ 77 > DETAILS EXPORT BIBTEX EXPORT RIS REFWORKS ENDNOTE EASYBIB CITATION PERMALIN LINKS GET IT View It CITATIONS Full Text Options ≥ > TAGS Details Title Development of a framework for knowledge translation: understanding user context Author Jacobson, Nora > Butterill, Dale > Goering, Paula > Health Services Research Subjects



Related reading recommended items that are related to the record



Demystifying knowledge translation: learning fro... Patricia Martens

Journal of health services research & policy.2005, Vol. 10(4), p. 203-11

ARTICLE suggested by



A guide to knowledge translation theory.

Carole A Estabrooks The Journal of continuing education in the health professions., Vol. 26(1), p. 25-36

ARTICLE suggested by

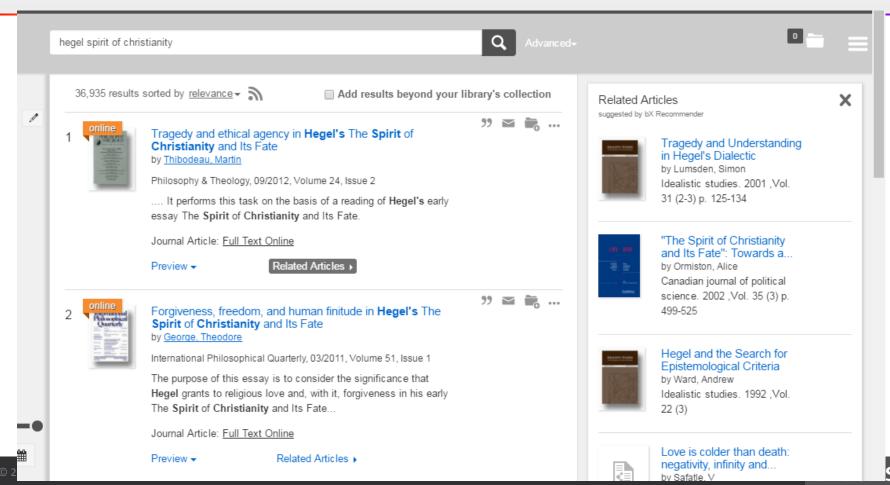


How can research organizations more... John N Lavis

The Milbank quarterly., Vol. 81(2), p. 221-48

ARTICLE SUggested by

bX Recommender on summon



Big data analytics – In the research lab



- 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

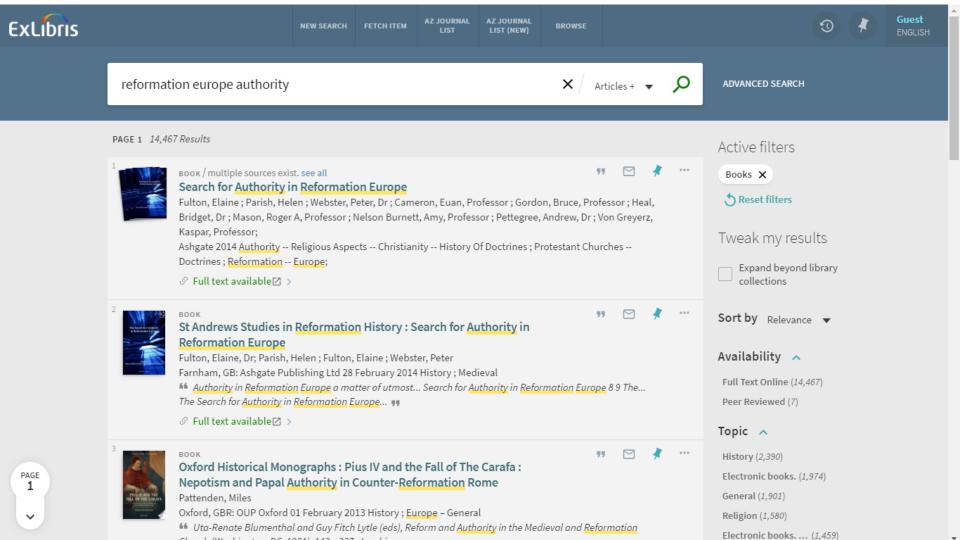
- Books and reviews
- Articles and research data
- •

Topic relations

...

• Topic exploration paths



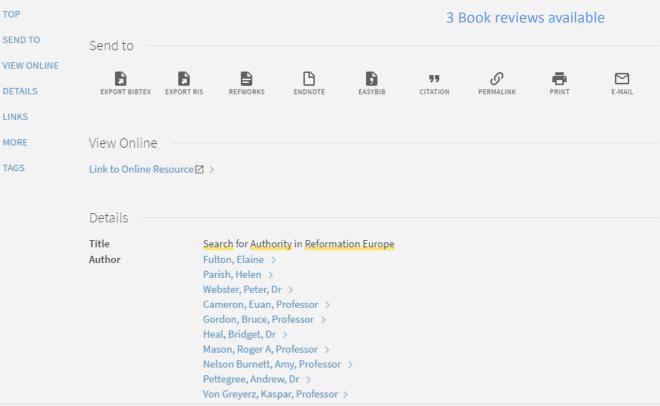


ExLibris

воок / multiple sources exist. see all

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; Ashgate 2014 <u>Authority</u> -- Religious Aspects -- Christianity -- History Of Doctrines ; Protestant Churches -- Doctrines ; <u>Reformation</u> -- <u>Europe</u>;







воок / multiple sources exist. see all

Search for Authority in Reformation Europe

NEW SEARCH

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;

AZ JOURNAL

LIST (NEW)

BROWSE

AZ JOURNAL

LIST

Guest

...

€

"

 REVIEW / multiple sources exist. see all The Search for Authority in Reformation Europe (Book Review) Rittgers, Ronald K. Visual Arts ; History & Archaeology ; Drama ; Languages & Literatures; Renaissance Quarterly, 01 March 2015, Vol.68(1), pp.341-343[Peer Reviewed Journal] Full text available > REVIEW / multiple sources exist. see all Helen Parish, Elaine Fulton and Peter Webster, eds, The Search for Authority in Reformation Europe Atherton, Ruth Parish, Helen ; Fulton, Elaine ; Webster, Peter; European History Quarterly, 2015, Vol.45(3), pp.589-591[Peer Reviewed Journal]
Helen Parish, Elaine Fulton and Peter Webster, eds, The Search for Authority in Reformation Europe Atherton, Ruth Parish, Helen ; Fulton, Elaine ; Webster, Peter;

Big data analytics – In the research lab



- 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





christine.stohn@exlibrisgroup.com dave.hovenden@exlibrisgroup.com eddie.neuwirth@exlibrisgroup.com

