

**OCCLC cataloging
community meeting**

The OCCLC cataloging community meeting

12 FEBRUARY 2025

#OCLCcataloging





David Whitehair

Director, Metadata & Digital Services
OCLC

Welcome back

Today's schedule



- Welcome
- Diversity, equity, and inclusion
- Open questions
- OCLC update 1
- Break
- OCLC update 2
- Open questions
- Closing

Deepening connections by turbocharging data with AI

OCCLC cataloging community meeting



**Bemal
Rajapatirana**

Director, WorldCat Data
Metadata & Digital Services
OCLC



Chelsea Dalgord

Product Manager,
Metadata & Digital Services
OCLC



Inkyung Choi

Associate Research Scientist
OCLC Research



Laura Ramsey

Senior Metadata Operations Manager,
WorldCat & KB Metadata Quality
OCLC



Latifa Baali

Access Services Librarian | Arabic
& Translation Studies Liaison
Librarian
American University of Sharjah

OCLC's AI: Services and Data



Library data



Library services

- Sustainable, efficient and responsible AI innovation
- Tailored to library values and community needs
- Enhancing library data
- AI + librarian expertise

Merging records with AI

Machine Learning Deduplication Project Update (1)

- Where we have been
 - First machine learning model ran in August 2023 eliminating over **5.4 million** duplicates
 - In the same timeframe, DDR eliminated **273,753** duplicates, and over **4.6** million the entire fiscal year
 - Model based on feedback of over 34,000 duplicates from over 300 users
 - Duplicates were eliminated for printed books in English, French, German, Italian, and Spanish

Machine Learning Deduplication Project Update (2)

- What we are wanting to achieve
 - Continuing to develop the model to leverage the potential of processing at scale to deduplicate WorldCat as a whole
 - Run our traditional Duplicate Detection and Resolution algorithms concurrently on new and updated records
 - Eliminate the straightforward duplicates and leave the complicated situations for the experts to resolve

Machine Learning Deduplication Project Update (3)

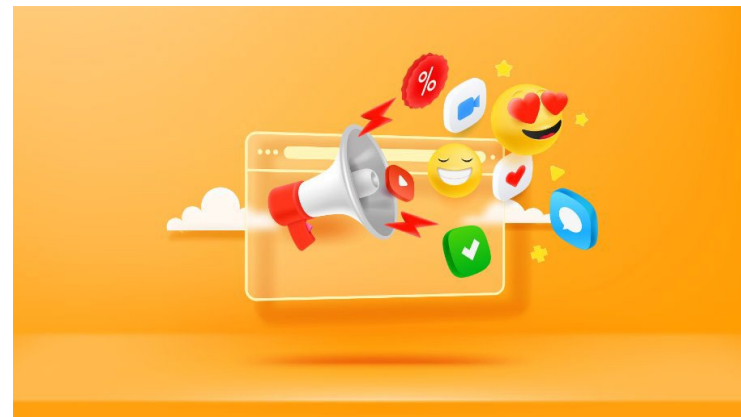
- Where we are now
 - Extending model to include all formats, languages, and scripts in WorldCat for version 2
 - Completed extensive internal testing
 - Adjusted model based on version 1 feedback and internal review of version 2
 - Engaged Member Merge libraries for verification of the algorithm

Member Merge Community Feedback (1)

- 17 libraries provided feedback on version 2 of the model
- Pairs of records included books, non-books, and non-English materials
 - Books, Continuing Resources, Maps, Scores, Sound Recordings, Visual Materials
 - Non-Latin scripts in Arabic, Bengali, Chinese, Cyrillic, Hebrew, Hindi, Japanese, Greek, Persian

Member Merge Community Feedback (2)

- Pairs included positive, negative, and questionable duplicates
- Further adjustments made based on member feedback



Example of positive pair

\$a De Janvry, Alain.	d100	Missing Entry
\$a Agricultural and rural development in Latin America : \$b new directions and new challenges / \$c by Alain de Janvry, Nigel Key and Elisabeth Sadoulet.	d245	\$a Agricultural and rural development policy in Latin America : \$b new directions and new challenges.
\$a Rome : \$b FAO, \$c 1997.	d260	\$a Rome : \$b FAO, \$c 1997.
\$a 56 pages : \$b illustrations ; \$c 30 cm.	d300	\$a 56 pages.
\$a text \$b txt \$2 rdacontent	d336	\$a text \$b txt \$2 rdacontent
\$a unmediated \$b n \$2 rdamedia	d337	\$a unmediated \$b n \$2 rdamedia
\$a volume \$b nc \$2 rdacarrier	d338	\$a volume \$b nc \$2 rdacarrier
\$a FAO agricultural policy and economic development series, \$x 1020-6531 ; \$v 2	d490	\$a FAO agricultural policy and economic development series ; \$v 2

Model, Metadata Quality, and Member Merge libraries all agreed

Example of negative pair

\$a 202009908X \$q (pbk.)	d020_0	\$a 2867440688
\$a 9782020099080 \$q (pbk.)	d020_1	\$a 9782867440687
\$a Sallenave, Danièle, \$d 1940-	d100	\$a Sallenave, Danièle, \$d 1940-
\$a La vie fantôme : \$b roman / \$c Danièle Sallenave.	d245	\$a La vie fantôme : \$b roman / \$c Danièle Sallenave.
\$a Paris : \$b P.O.L., \$c ©1986.	d260	\$a Paris : \$b P.O.L., \$c ©1986.
\$a 284 pages ; \$c 18 cm	d300	\$a 288 pages ; \$c 21 cm
\$a text \$b txt \$2 rdacontent	d336	\$a text \$b txt \$2 rdacontent
\$a unmediated \$b n \$2 rdamedia	d337	\$a unmediated \$b n \$2 rdamedia
\$a volume \$b nc \$2 rdacarrier	d338	\$a volume \$b nc \$2 rdacarrier
\$a Points ; \$v R299	d490	Missing Entry

Model, Metadata Quality, and Member Merge libraries all agreed

Example of questionable pair

\$a Œil.	d130	Missing Entry
\$a Aspects of modern art : \$b an anthology of writings on modern art from L'Œil, the European art magazine / \$c edited by Georges and Rosamund Bernier ; with 40 pages in color.	d245	\$a Aspects of modern art : \$b an anthology of writings on modern art from l'Oeil, the European art magazine / \$c edited by Georges and Rosamond Bernier ; with 40 pages in color.
\$a London : \$b A. Zwemmer ; \$a Paris : \$b G. & R. Bernier, \$c [1957]	d260	\$a Lausanne : \$b Bernier ; \$a London : \$b Zwemmer, \$c [1957]
\$a 188 pages : \$b illustrations (some color) ; \$c 32 cm.	d300	\$a 188 pages : \$b illustrations (some color) ; \$c 32 cm.
\$a text \$b txt \$2 rdacontent	d336	\$a text \$b txt \$2 rdacontent
\$a unmediated \$b n \$2 rdamedia	d337	\$a unmediated \$b n \$2 rdamedia
\$a volume \$b nc \$2 rdacarrier	d338	\$a volume \$b nc \$2 rdacarrier

Mixed responses between the model, Metadata Quality, and Member Libraries

Challenges—sparse data

- Record 1

```
100 1 Silverstein, Shel.  
245 1 2 A Giraffe and A Half.  
260      ⚭c 1964.
```

- Record 2

```
100 1 Silverstein, Shel.  
245 1 2 A giraffe and a half.  
260      ⚭b Evil Eye Music, Inc. ⚭c ©1964.  
300      44 pages
```


Challenges—slight variances that matter

```
100 1    Hof, J. van 't, ꞑc schrijfonderwijs. ꞑ0 (NL-LeOCL)30597677X
245 1 0  Van 't Hof's Schrijfcursus in 19 nummers. ꞑn VIIA / ꞑc J. van 't Hof.
```

```
100 1    Hof, J. van 't, ꞑc schrijfonderwijs. ꞑ0 (NL-LeOCL)30597677X
245 1 0  Van 't Hof's Schrijfcursus in 19 nummers. ꞑn VIIIA / ꞑc J. van 't Hof.
```

Note: Fields 260 and 300 are the same on these records

Challenges—confidence in the process (1)

“Machine learning doesn't work in this case”

```
110 2    New Brunswick Electric Power Commission.  
245 1 0  1984 / ꞑc 85 annual report.  
260      Fredericton : ꞑb The Power Commission, ꞑc 1985.  
300      32 pages : ꞑb illustrations
```

Note: These records won't be merged!

```
100 1    Divaris, C.  
245 1 0  1984 / ꞑc 85 Old mutual income tax guide.  
260      ꞑc 1985.
```

Challenges—confidence in the process (2)

“Duplicate, but human eyes required to verify”

“These are probably the same thing, but I would not merge due to differences in 260 \$b”

“Lacking too much info for me to be confident they are same thing, especially since this is a popular work that has been published a zillion times”

Solutions to the challenges

- Taking a more conservative approach
- Eliminating records where slight differences may matter
- Validating model against record pairs from version 1
- Future—collecting more data from the community to continue to train the model

Next steps

- Process 500,000 record pairs of print book records—started running on Feb. 11
- Evaluate member feedback and success of the run before resuming
- Process an additional 2.25 million pairs of print book records over the course of several weeks

How can you help?

- Report suspicious merges to bibchange@oclc.org
- Participate in the next data labelling project
 - Look for announcements coming in the next few weeks
 - Non-books and non-Latin expertise is welcome
- Send questions to ASKQC@oclc.org

OCLC Member Merge Program & De- duplication Project: Personal experience reviewing Arabic duplicates

Latifa Baali

Access Services Librarian
American University of Sharjah
Email: LBAALI@AUS.EDU

February 12th, 2025



Member Merge Program: Arabic Records

How did I learn about the Member Merge program?

- **May 2023:** A message was sent to MELANET-L from Iman Dagher, the Arabic & Islamic Studies Metadata Librarian at the UCLA Library, inviting Arabic script catalogers to join the OCLC Member Merge Program (for the purpose of evaluating and merging duplicate records in OCLC).
- **Participants in the OCLC Member Merge Program – Arabic Records:**
 - Institutions inside the U.S.: American Islamic College, Northwestern University, Penn State University, UCLA, and University of Nevada.
 - Institutions outside the U.S.: American University of Sharjah, American University of Cairo and Library of Congress – Cairo Office.
- **June 2023:** First meeting with the OCLC reviewer Shanna Griffith, OCLC Senior Data Analyst, Data Quality & Governance, Global Technologies.

Member Merge program : training & practice

- Five one-hour long webinars were conducted over several weeks.
- Participants received comprehensive instructional documentation and tutorial videos.
- Common exercises which included hands-on practice related to applying training materials and guidelines to make sound decisions for merging records.
- Connexion client authorization was upgraded to allow participant to merge records.
- Six batches of over 30 sets of records (three sets per member); plus 100 sets of records (mostly Arabic and a few were Persian, English and French records) were reviewed by participants and submitted to Shanna Griffith for feedback and merge approval.

CERTIFICATE *of* ACHIEVEMENT

THIS ACKNOWLEDGES THAT

Latifa Baali

HAS GAINED INDEPENDENCE IN MERGING

Print and Electronic Monographs
in the
Member Merge Program

FEBRUARY 14,
2024

Shanna Griffith

Reviewer

Laura Ramsey

Senior Metadata Operations Manager, WorldCat and KB
Metadata Quality




Latifa Baali

Arabic Records Merging FY 2024-2025

	Month	Original number of records	Number of records after merging
2024	July	8	4
	August	11	5
	September	6	3
	October	61	29
	November	24	10
	December	56	28
2025	January	12	6
Total		178	85

October 2023 : A message from Cynthia M. Whitacre, OCLC Senior Metadata Operations Manager, inviting Metadata experts in non-Latin languages to participate in a data labeling exercise aiming at validating OCLC machine learning model's understanding of duplicates.

November – December 2023 : I have reviewed and labeled **100** sets of pair records (Arabic print and electronic books) using OCLC data labeling tool.



Are these records analogous?

For questions or assistance contact us at datalabeling@oclc.org

[Login](#)
[Leaderboard](#)
[FAQ](#)

Random Pair

Generate new pair

Select language of resource

ARA

Select material type

Book

Select age of record

After 1830

Generate Pair

Do these two records describe the same item?

☐ Yes
☐ No
☐ Unsure

Use checkboxes to select the fields most relevant to your decision

Submit Decision

October 2024: a message was sent from Laura Ramsey, to the Member Merge Participants to test a small set of record pairs (less than 50 sets) generated through test runs with most recent algorithms.

I have reviewed **25** sets of Arabic records of print and electronic monographs:

Criteria used to compare between records in each pair:

- Language of cataloging.
- Item format.
- Item language.
- Bibliographic information in MARC fields 020, 100, 245, 250, 260/264, 300.

Review Result : the 25 sets were duplicates!

(Pinned) Online WorldCat: Azhar al-Mawt

OCCL 1296392028 No holdings in A7U - 1 other holding

Books	Rec stat	c	Entered	20201001	Replaced	202
Type a	ELvl M	Srcd	Audn	Ctrl	Lang ara	
BLvl m	Form	Conf 0	Biog	MRec	Ctry le	
	Cont	GPub	LitF 1	Indx 0		
Desc i	Ills	Fest 0	DtSt s	Dates 2020		

040 AU@ #b eng #e rda #c AU@ #d OCLCO #d OCLCF #d OCLCQ

066 #c (3)

020 9786144850442 #q (paperback)

020 6144850449 #q (paperback)

082 0 4 892.73 #q OCoLC #2 23/eng/20230216

090 #b

049 A7UA

100 1 غازی، فایز، #e author.

100 1 Ghāzi, Fāyiz, #e author.

245 1 0 ازهار الموت : #b رواية / #c فایز غازی.

245 1 0 Azhar al-Mawt : #b riwāyah / #c Fāyiz Ghāzi.

264 1 Bayrūt : #b Dār al-Fārābi, #c 2020.

300 135 pages ; #c 21 cm.

336 text #b txt #2 rdacontent

(Pinned) Online WorldCat: Azhar al-Mawt

OCCL 1240111841 No holdings in A7U - 2 other holdings

Books	Rec stat	c	Entered	20201001	Replaced	202
Type a	ELvl M	Srcd	Audn	Ctrl	Lang ara	
BLvl m	Form	Conf 0	Biog	MRec	Ctry le	
	Cont	GPub	LitF 1	Indx 0		
Desc i	Ills	Fest 0	DtSt s	Dates 2020		

040 NZAUC #b eng #e rda #c NZAUC #d OCLCO #d OCLCF #d LEILA #d OCLCQ

066 #c (3)

020 9786144850442 #q (paperback)

020 6144850449

082 0 4 892.73 #q OCoLC #2 23/eng/20231120

090 #b

049 A7UA

100 1 غازی، فایز، #e author.

100 1 Ghāzi, Fāyiz, #e author.

242 1 0 Blossoms of death. #y eng.

245 1 0 ازهار الموت : #b رواية / #c فایز غازی.

245 1 0 Azhar al-Mawt : #b riwāyah / #c Fāyiz Ghāzi.

264 1 Bayrūt : #b Dār al-Fārābi, #c 2020.

300 135 pages ; #c 21 cm.

I believe that employing machine learning technology in the OCLC De-duplication project is an effective strategy for reducing duplicate records. I anticipate that the implementation of this technology will yield the following benefits:

- ↳ Enhancement in the quality and accuracy of bibliographic records.
- ↳ Increase of the discoverability of library bibliographic data.
- ↳ Time-saving for catalogers and Interlibrary Loan staff worldwide.

In my opinion, this is a good example of a useful application of Artificial Intelligence!

Enriching records with AI

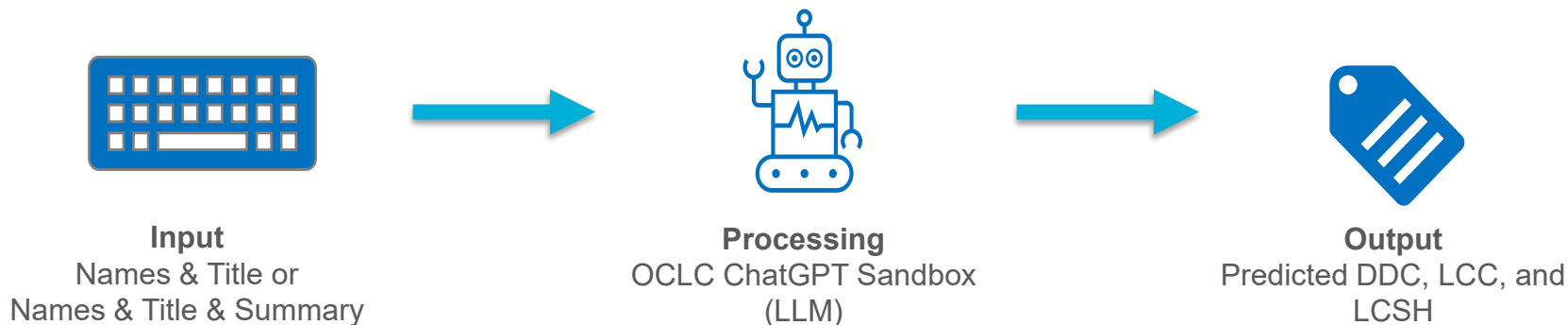
Enriching records with AI for better discoverability

- Subject analysis accounts for 70-80% of the time to catalog a record
- Average time to catalog a single record:
 - Copy cataloging: average 20 minutes
 - Original cataloging: average 40 minutes
- Predicting subjects and classifications



Experiment 1:

Can ChatGPT accurately predict subject headings and classifications?



FINDINGS:

- Untrained LLM is correct only 60% of the time
- Invalid answers (hallucinations) are a challenge

NEXT STEP:

- Leverage data ecosystem for context and creditability

Experiment 2:

Do results get more accurate if we add WorldCat data as context?



FINDINGS:

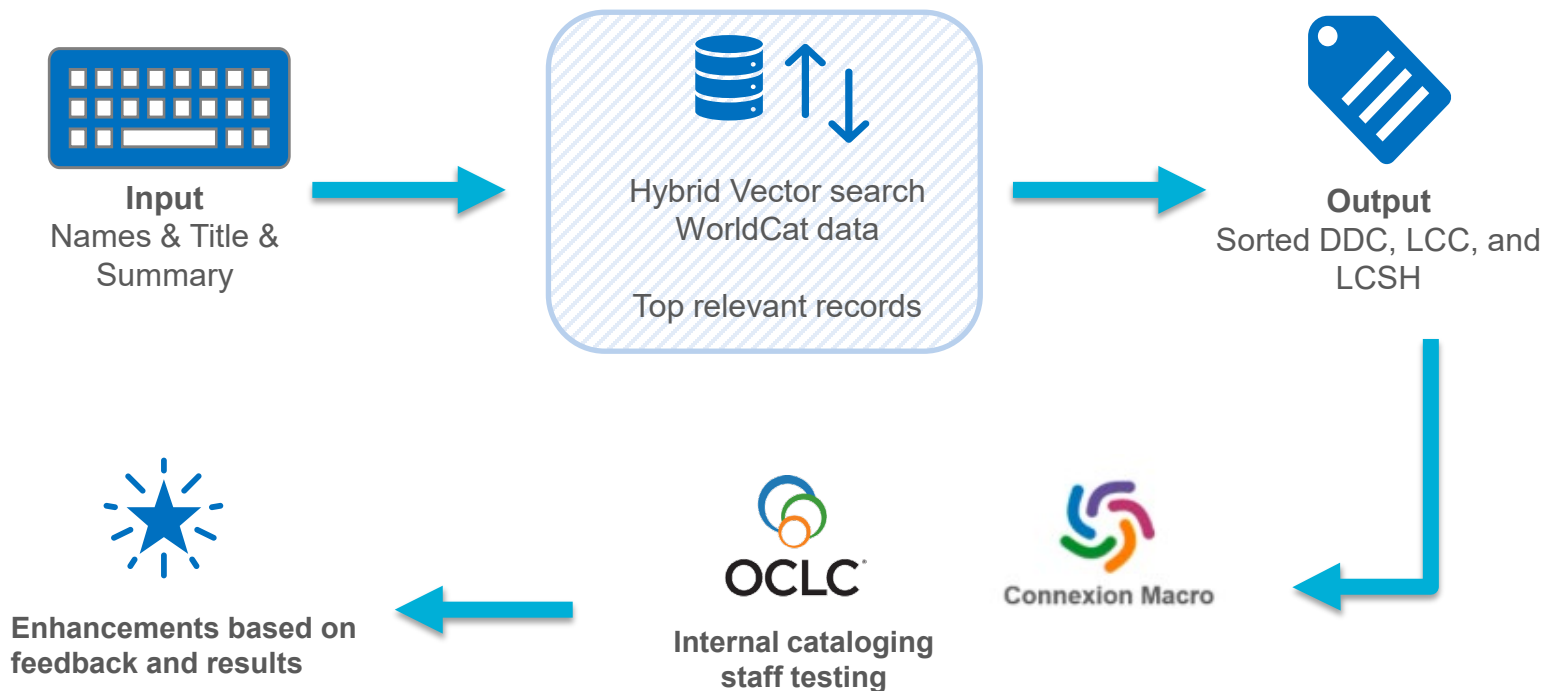
- Response times for generating relevant records for cataloging workflow was not feasible
- Ranked results are very promising!

NEXT STEP:

- Replace vector store with Cortex index search
- Human feedback loop

Experiment 3 & 4

Can we improve query speed and relevance while keeping or improving accuracy?



020		0688060692 #q (pbk.)
020		9780688060695 #q (pbk.)
043		n-us---
050	0 0	CT275.P648 #b A3 1984 Guidebook
082	0 0	917.304/92 #2 20
096		917.3 P672zd
090		#b
049		ZMZA
100	1	DiSanto, Ronald L.
245	1 0	Guidebook to Zen and the art of motorcycle maintenance / #c Ronald L. DiSanto.
250		1st ed.
260		New York : #b W. Morrow, #c ©1990.
300		407 pages : #b illustrations, map : #c 25 cm
336		text #b txt #2 rdacontent
337		unmediated #b n #2 rdamedia
338		volume #b nc #2 rdacarrier
504		Includes bibliographical references (pages 338-357) and index.
520		When Robert Pirsig's Zen and the Art of Motorcycle Maintenance was first published, it was a sensation. It was the first book to introduce the general public to the ideas of Zen Buddhism, and its author, Robert Pirsig, was a leading authority on the subject. In this new edition, Pirsig has added a new introduction, a new chapter on the history of Zen, and a new chapter on the history of the book. This new edition is a must-read for anyone interested in Zen, Buddhism, or the history of the book.
600	1 0	Pirsig, Robert M. #t Zen and the art of motorcycle maintenance.
600	1 0	Pirsig, Robert M.
600	1 0	Pirsig, Robert M. #x Travel #z United States.
650	0	Fathers and sons #z United States.
650	0	Philosophy and civilization.

OCCLC AI Enrichment

DDC fields to choose from:

- ☐ 917.304/92 #2 20
- ☐ 111.1 #2 20
- ☐ 973.09/92 #a [B] #2 23
- ☐ 294.3
- ☐ 171 #a B #2 21

LCC fields to choose from:

- ☐ CT275.P648
- ☐ BJ1547.5.P648
- ☐ CT275.P648R53 2008

LCSH fields to choose from:

- ☒ Philosophy and civilization.
- ☒ Zen Buddhism and science.
- ☒ Essentialism (Philosophy)
- ☒ Values.
- ☒ Self.
- ☒ Conduct of life.
- ☒ Ego (Psychology)
- ☐ Fathers and sons
- ☐ Voyages and travels.
- ☐ Zen Buddhism.

NOTE: Fields that are both selected and disabled are already in the record.

OK
Cancel

Quotes from OCLC cataloging staff

*"I do think that it could **benefit catalogers who have limited cataloging time.**"*

*"From what I observed as a tester, I have great confidence in the potential of this AI model to **enhance cataloging efficiency**, so that libraries can better serve their patrons."*

*"Overall, I think it will be a **time-saver** and I miss not being able to use it now."*

Next steps

- Complete second round of testing with internal cataloging staff at OCLC
- Pilot with libraries using Connexion this quarter



The power of e-holdings automations

OCCLC cataloging community meeting



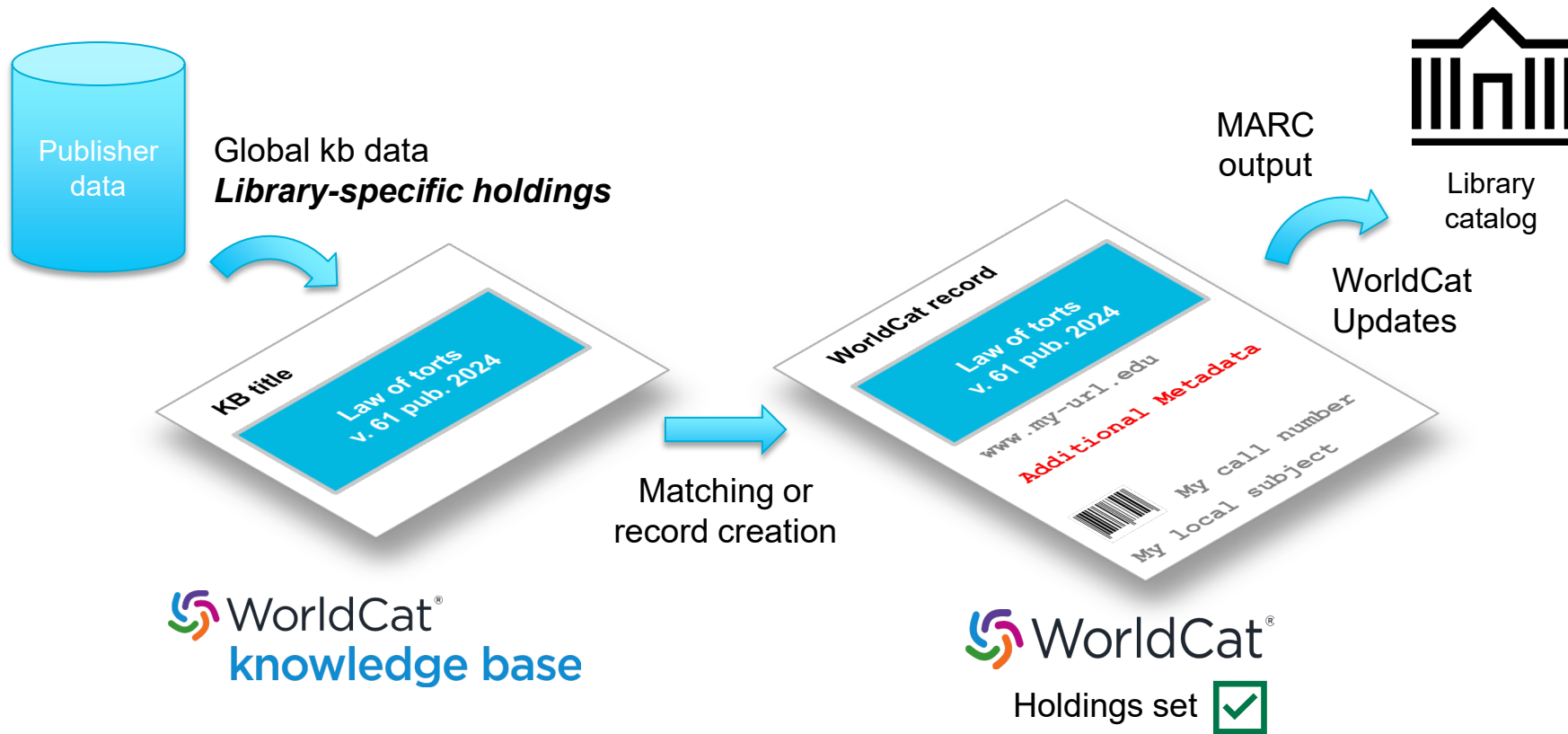
Hank Sway

Product Manager,
Metadata & Digital Services
OCCLC



Dmitrijs Martinovs

Product Manager,
Sage



Key benefits

Current

- No manual title selection!
- Important for collections purchased title by title
- Holdings automatically set or removed in WorldCat as your collection changes
- MARC record delivery

Future

- Additional providers
- Dashboard for automatic holdings management

WorldCat knowledge base collections

Automate your holdings management



EBSCO

ELSEVIER



Knovel®

Ovid®



Rittenhouse
BOOK DISTRIBUTORS



SPRINGER NATURE



Taylor & Francis Group
an informa business



WILEY

Future partners:



CloudLibrary



#OCLCcataloging

oclc.org/autoload





- Manual process
- Time spent identifying packages
- Time spent modifying packages when subscriptions change
- Discoverability issues
- Lower usage

- Reduction in manual work
- Packages are activated automatically
- Automation keeps holdings up to date
- Improved discoverability
- Higher usage

Sage Research Methods

**Supercharging
research**

- Data and Research Literacy
- Research Design and Planning (Video)

Sage Skills

**Building
confidence**

- Student Success Part 2

With researchers. With integrity. **With impact.**

Sage Journals

Sage Video

**Streaming
knowledge**

- Disability Studies

Sage Business

**Shaping
futures**

- Business Foundations Part 2

Sage Reference & Academic Books

**Broadening
minds**

Supporting materials to get started

[KBART automation
implementation Guide
\(Sage Journals\)](#)

[KBART automation
implementation Guide
\(Learning Resources\)](#)

[Automation FAQs
\(Sage Journals\)](#)

[Automation FAQs
\(Learning Resources\)](#)

Sage Journals

Discover
world-class
research

Discoverability Checklist

Let us help you and your patrons get the most out of Sage Journals. Here are our top tips for increasing the discoverability of your important content.

Adding Sage Journals products to your catalog and library websites

- ✓ List all the Sage Journals products in your A-Z and/or subject directory of databases
- ✓ Use the recommended product description
We recommend cataloging Sage Journals as eJournals by individual title – not as a database.
 - Won more than 75% of Sage's journals ranked in the Journal Citation Reports (Source: Clarivate, 2024). Sage Journals is host to high-impact, peer-reviewed, and interdisciplinary articles – the heart of your research. Now on a new, user-friendly platform, journals and articles can be searched, viewed, and shared with ease from any device. Article-level metrics – including social media sharing – from Altmetrics and Journal-level Impact Factor rankings help students, faculty, and researchers understand the reach and importance of the articles they are doing.
 - The Sage Journals portfolio includes more than 1,100 journals spanning the humanities, social sciences, and science, technology, engineering, and medicine. Whether your library needs the full, interdisciplinary collection dating back to volume 1, issue 1, or a specific, focused collection targeted to your researchers' interests, Sage can help you build the perfect fit.
- ✓ Appoint a "super user"
 - Help your patrons onboard to Sage Journals by offering resource training.
 - Meet our training team.
- ✓ Display our Sage Journals poster in your library
 - View all of our downloadable usage-driving materials.

Finding Sage Journals through discovery services

Sage Journals metadata is regularly delivered to e-resource management services, such as:	Sage Journals are fully indexed by third-party discovery products and services, including:
Central (NCBI/NLM)	bing
ERIC (U.S. Department of Education)	EBSCO Discovery Service
PsycINFO (American Psychological Association)	Ex Libris Primo
PubMed (MEDLINE and PubMed)	Google
Scopus (Elsevier)	Google Scholar
Web of Science (Clarivate Analytics)	OCUL World Cat Discovery
	ProQuest Summon
	Yahoo!

Set-up KBART files
Accurately reflect institutional entitlements, including Sage Gold OA titles, in your discovery service.

For questions or suggestions about the discoverability of Sage Journals in your library, please contact us at:
librarymktg@sagepub.com

journals.sagepub.com

KBART Automation: Prioritize Expertise, Minimize Repetition



OCLC cataloging update



Webex:

Please submit questions in chat
and address the chat to “everyone”

Vimeo livestream:

Please use the form on the
livestream page to submit
questions

Final notes

Recordings, slides & links

- Watch your email
 - Webex email
 - Listservs
- OCLC will post links to content in the [Cataloging & Metadata](#) community
- Series site: oclc.org/cataloging-community-meetings

Final notes

Recordings, slides & links

- Watch your email
 - Webex email
 - Listservs
- OCLC will post links to content in the [Cataloging & Metadata community](#)
- Series site: oclc.org/cataloging-community-meetings

Continue the conversation

- Cataloging policy
 - Email AskQC@oclc.org
 - AskQC twice monthly webinars
- News, discussions, and more...
 - Cataloging and Metadata [community](#)
 - Record Manager [community](#)
 - Collection Manager [community](#)
 - Meridian [community](#)

**OCCLC cataloging
community meeting**

thank you

Stay in touch

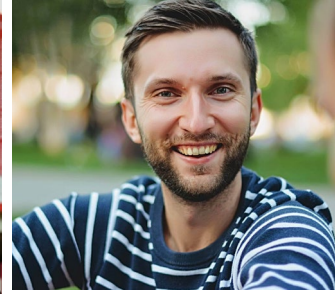
Regional offices:

oc.lc/offices

Community Center:

oc.lc/community

Thank you



**OCCLC cataloging
community meeting**

thank you

**And thank you to
our community!**

**OCCLC cataloging
community meeting**

thank you

12 February 2025

#OCLCcataloging

