Virtual AskQC Office Hours

Data and algorithms and bibs, oh my!

OCLC Metadata Quality

April 2023

OCLC

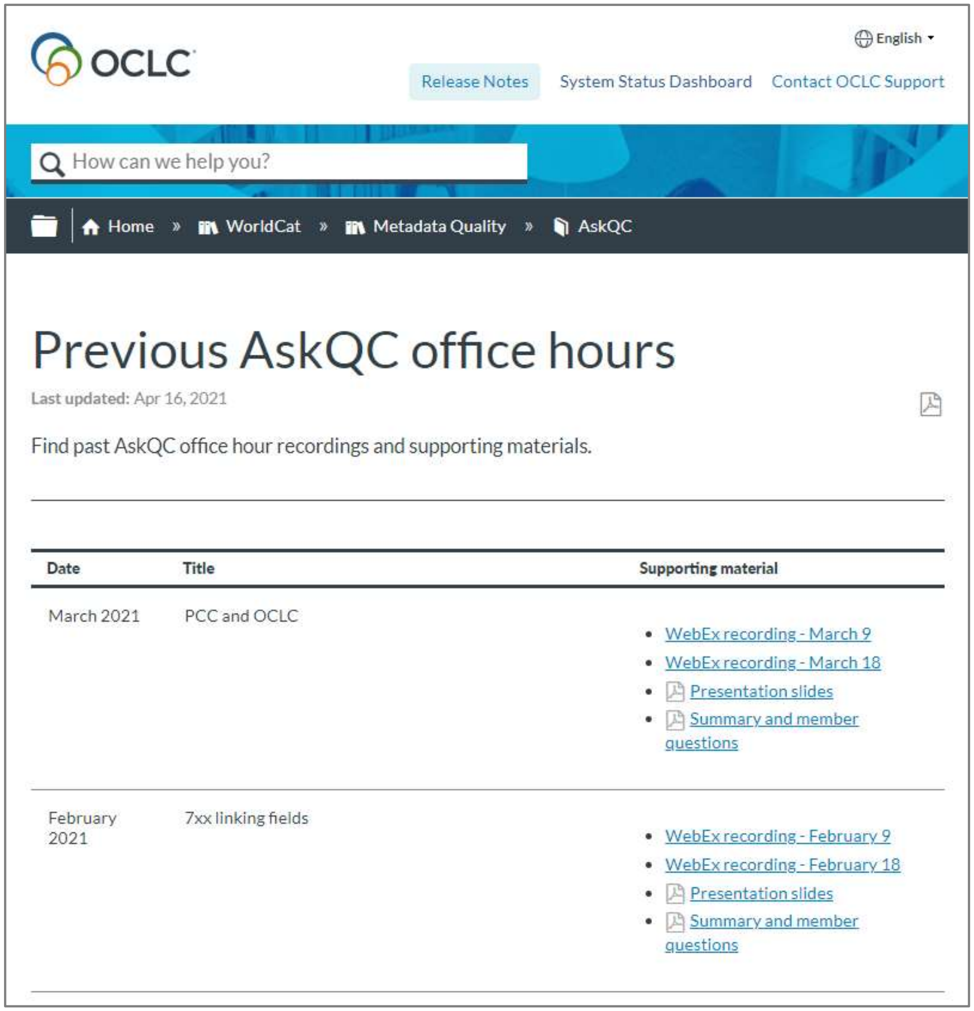
Housekeeping

This session is being recorded

Housekeeping

This session is being recorded

All session recordings, slides, and notes are available at oc.lc/askqc



Previous AskQC office hours

Last updated: Apr 16, 2021

Find past AskQC office hour recordings and supporting materials.

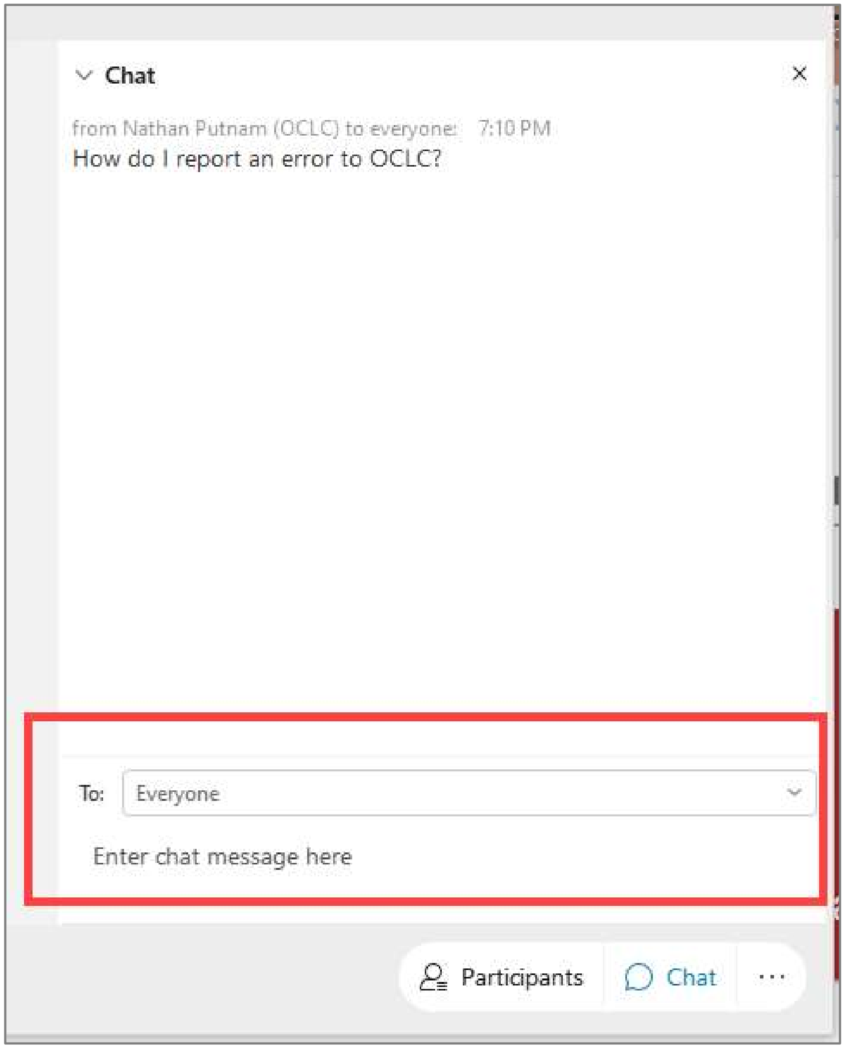
Date	Title	Supporting material
March 2021	PCC and OCLC	- WebEx recording - March 9 - WebEx recording - March 18 - Presentation slides - Summary and member questions
February 2021	7xx linking fields	- WebEx recording - February 9 - WebEx recording - February 18 - Presentation slides - Summary and member questions

Housekeeping

This session is being recorded

All session recordings, slides, and notes are available at oc.lc/askqc

Enter questions in chat to "Everyone" at any time during the presentation



from Nathan Putnam (OCLC) to everyone: 7:10 PM
How do I report an error to OCLC?

To: Everyone

Enter chat message here

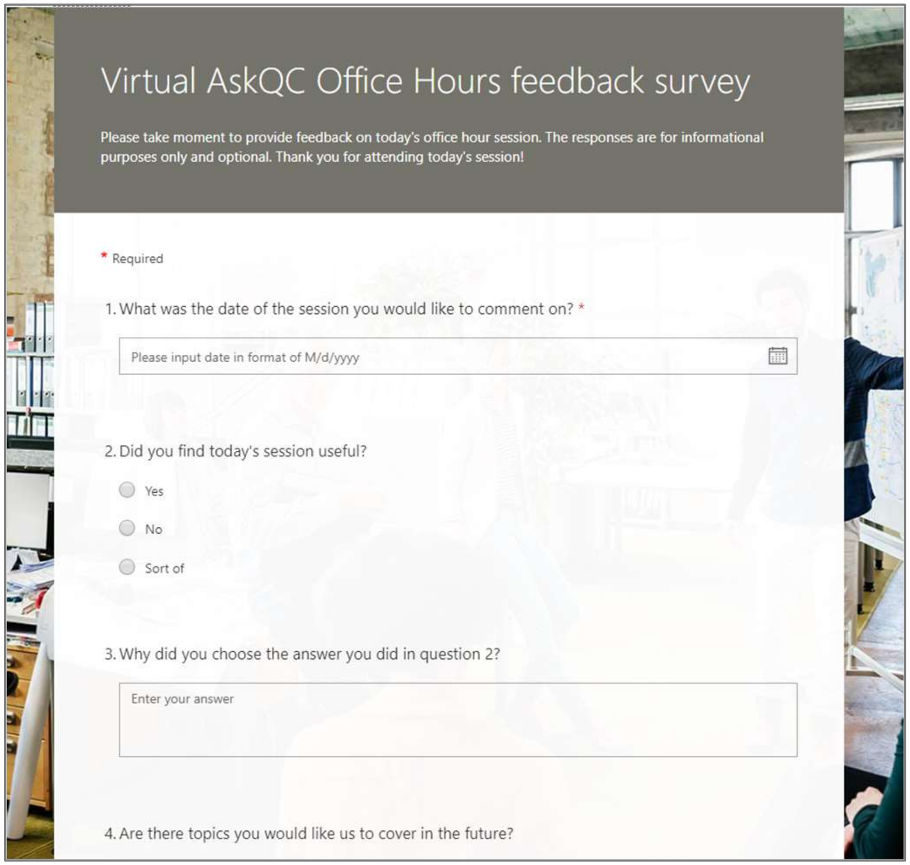
Housekeeping

This session is being recorded

All session recordings, slides, and notes are available at oc.lc/askqc

Enter questions in chat to "Everyone" at any time during the presentation.

After the session, you will be directed to a brief, optional survey



Virtual AskQC Office Hours feedback survey

Please take moment to provide feedback on today's office hour session. The responses are for informational purposes only and optional. Thank you for attending today's session!

* Required

1. What was the date of the session you would like to comment on? *

Please input date in format of M/d/yyyy

2. Did you find today's session useful?

○ Yes

○ No

○ Sort of

3. Why did you choose the answer you did in question 2?

Enter your answer

4. Are there topics you would like us to cover in the future?

On the call today

Hayley Moreno

Senior Data Analyst

Laura Ramsey

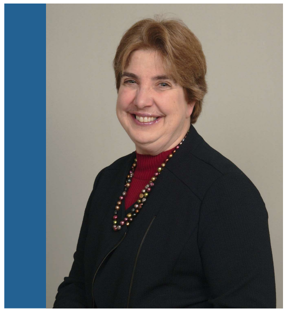
Senior Metadata Operations Manager

Nathan Putnam

Director, Data Quality & Governance

Jenny Toves

Senior Technical Manager



Becky Dean

Lead Data Analyst

Robin Six

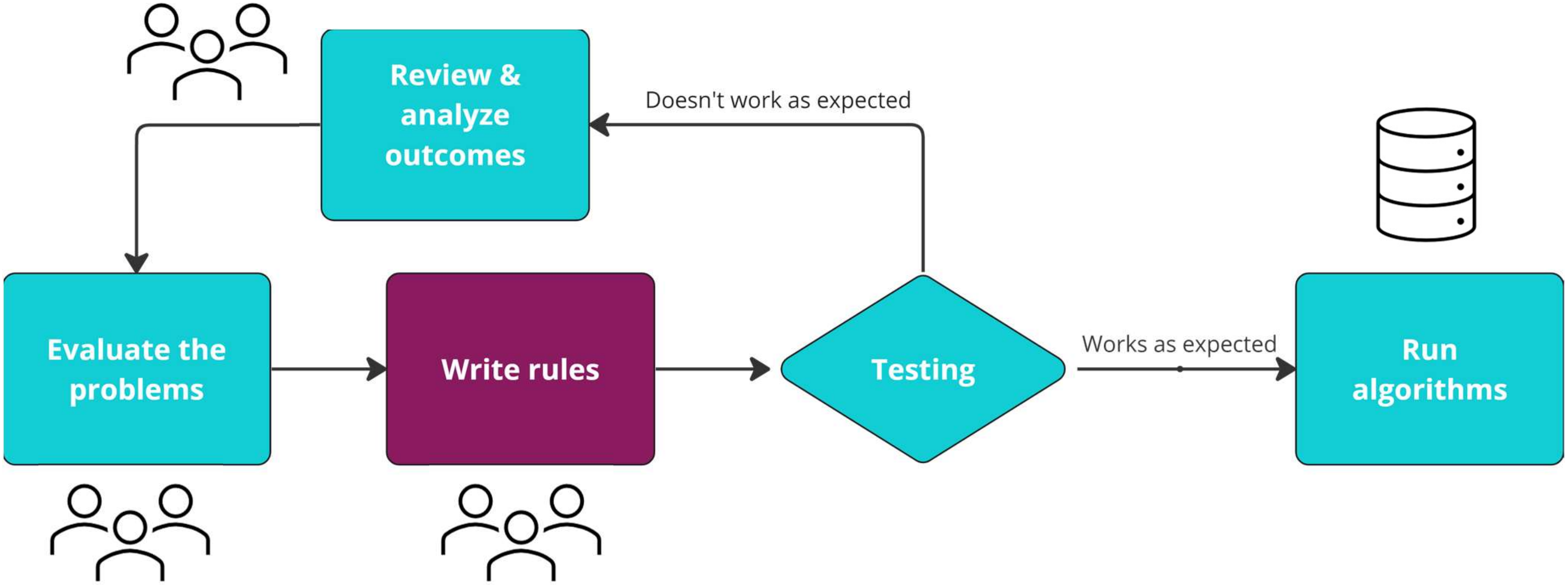
Senior Data Analyst

Data and algorithms and bibs, oh my!

Laura Ramsey
Senior Metadata
Operations Manager

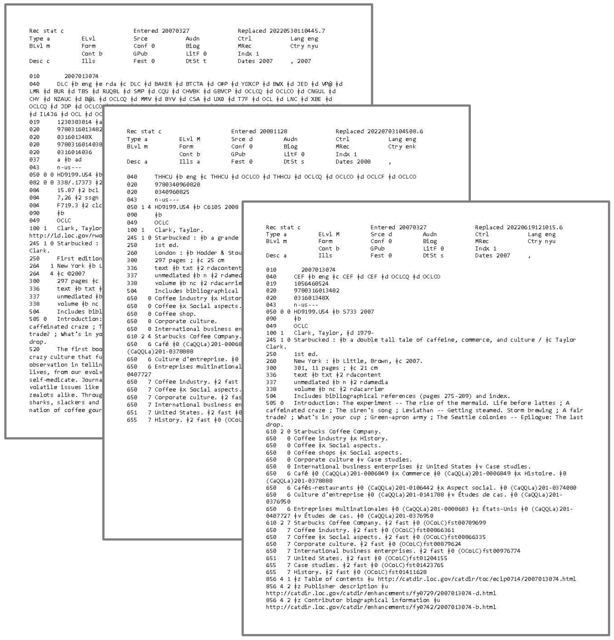
Nathan Putnam
Director,
Data Quality & Governance

Rules-based algorithms



Adapted from: Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. Second edition. O'Reilly Media. Kindle Edition.

Rules-based algorithms

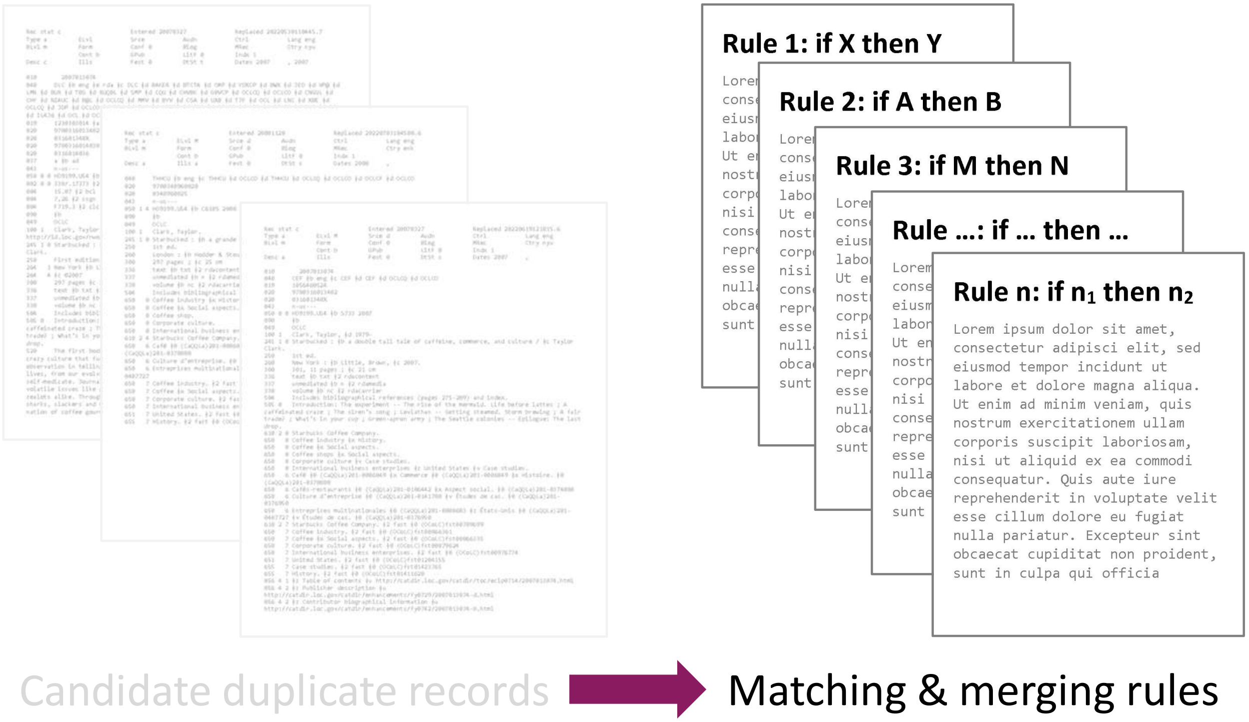


```
Rec stat c                    Entered 20070327              Replaced 20220619121015.6
Type a          ELvl M        Srce d        Audn            Ctrl          Lang eng
BLvl m          Form          Conf 0        Biog            MRec          Ctry nyu
                Cont b        GPub          LitF 0          Indx 1
Desc a          Ills          Fest 0        DtSt s          Dates 2007    ,

010       2007013074
040     CEF ǂb eng ǂc CEF ǂd CEF ǂd OCLCQ ǂd OCLCO
019     1056468524
020     9780316013482
020     031601348X
043     n-us---
050 0 0 HD9199.U54 ǂb S733 2007
090     ǂb
049     OCLC
100 1   Clark, Taylor, ǂd 1979-
245 1 0 Starbucked : ǂb a double tall tale of caffeine, commerce, and culture / ǂc Taylor
Clark.
250     1st ed.
260     New York : ǂb Little, Brown, ǂc 2007.
300     301, 11 pages ; ǂc 21 cm
336     text ǂb txt ǂ2 rdacontent
337     unmediated ǂb n ǂ2 rdamedia
338     volume ǂb nc ǂ2 rdacarrier
504     Includes bibliographical references (pages 275-289) and index.
505 0   Introduction: The experiment -- The rise of the mermaid. Life before lattes ; A
caffeinated craze ; The siren's song ; Leviathan -- Getting steamed. Storm brewing ; A fair
trade? ; What's in your cup ; Green-apron army ; The Seattle colonies -- Epilogue: The last
drop.
610 2 0 Starbucks Coffee Company.
650   0 Coffee industry ǂx History.
650   0 Coffee ǂx Social aspects.
650   0 Coffee shops ǂx Social aspects.
650   0 Corporate culture ǂv Case studies.
650   0 International business enterprises ǂz United States ǂv Case studies.
650   6 Café ǂ0 (CaQQLa)201-0006849 ǂx Commerce ǂ0 (CaQQLa)201-0006849 ǂx Histoire. ǂ0
(CaQQLa)201-0378888
650   6 Cafés-restaurants ǂ0 (CaQQLa)201-0106442 ǂx Aspect social. ǂ0 (CaQQLa)201-0374080
650   6 Culture d'entreprise ǂ0 (CaQQLa)201-0141788 ǂv Études de cas. ǂ0 (CaQQLa)201-
0376950
650   6 Entreprises multinationales ǂ0 (CaQQLa)201-0008683 ǂz États-Unis ǂ0 (CaQQLa)201-
0407727 ǂv Études de cas. ǂ0 (CaQQLa)201-0376950
610 2 7 Starbucks Coffee Company. ǂ2 fast ǂ0 (OCoLC)fst00709699
650   7 Coffee industry. ǂ2 fast ǂ0 (OCoLC)fst00866361
650   7 Coffee ǂx Social aspects. ǂ2 fast ǂ0 (OCoLC)fst00866335
650   7 Corporate culture. ǂ2 fast ǂ0 (OCoLC)fst00879624
650   7 International business enterprises. ǂ2 fast ǂ0 (OCoLC)fst00976774
651   7 United States. ǂ2 fast ǂ0 (OCoLC)fst01204155
655   7 Case studies. ǂ2 fast ǂ0 (OCoLC)fst01423765
655   7 History. ǂ2 fast ǂ0 (OCoLC)fst01411628
856 4 1 ǂz Table of contents ǂu http://catdir.loc.gov/catdir/toc/ecip0714/2007013074.html
856 4 2 ǂz Publisher description ǂu
http://catdir.loc.gov/catdir/enhancements/fy0729/2007013074-d.html
856 4 2 ǂz Contributor biographical information ǂu
http://catdir.loc.gov/catdir/enhancements/fy0742/2007013074-b.html
```

Candidate duplicate records

Rules-based algorithms



Rule 1: if X then Y

Rule 2: if A then B

Rule 3: if M then N

Rule …: if … then …

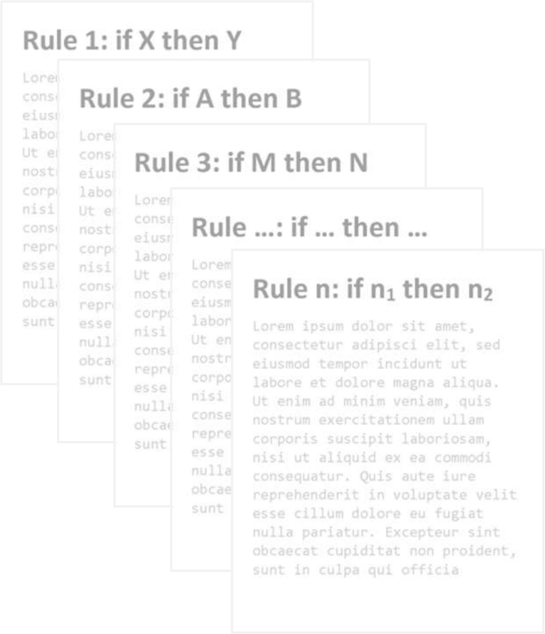
Rule n: if n_1 then n_2

Lorem ipsum dolor sit amet, consectetur adipisci elit, sed eiusmod tempor incidunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur. Quis aute iure reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint obcaecat cupiditat non proident, sunt in culpa qui officia

Candidate duplicate records → Matching & merging rules

Rules-based algorithms

Candidate duplicate records →



Rule 1: if X then Y

Rule 2: if A then B

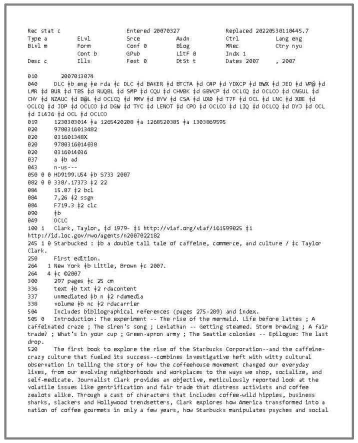
Rule 3: if M then N

Rule ...: if ... then ...

Rule n: if n_1 then n_2

Lorem ipsum dolor sit amet, consectetur adipisci elit, sed eiusmod tempor incidunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur. Quis aute iure reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint obcaecat cupiditat non proident, sunt in culpa qui officia

Matching & merging rules →



```
Rec stat c                    Entered 20070327              Replaced 20220530110445.7
Type a         ELvl           Srce           Audn           Ctrl           Lang eng
BLvl m         Form           Conf 0         Biog           MRec           Ctry nyu
               Cont b         GPub           LitF 0         Indx 1
Desc c         Ills           Fest 0         DtSt t         Dates 2007     , 2007

010       2007013074
040     DLC ǂb eng ǂe rda ǂc DLC ǂd BAKER ǂd BTCTA ǂd C#P ǂd YDXCP ǂd BWX ǂd JED ǂd VP@ ǂd
LMR ǂd BUR ǂd TBS ǂd RUQBL ǂd SMP ǂd CQU ǂd CHVBK ǂd GBVCP ǂd OCLCQ ǂd OCLCO ǂd CNGUL ǂd
CHY ǂd NZAUC ǂd B@L ǂd OCLCQ ǂd MMV ǂd BYV ǂd CSA ǂd UX0 ǂd T7F ǂd OCL ǂd LNC ǂd XBE ǂd
OCLCQ ǂd JDP ǂd OCLCO ǂd DGW ǂd TYC ǂd LENOT ǂd CPO ǂd OCLCO ǂd LIQ ǂd OCLCQ ǂd DYJ ǂd OCL
ǂd IL4J6 ǂd OCL ǂd OCLCO
019     1230303014 ǂa 1265420208 ǂa 1268520385 ǂa 1303869595
020     9780316013482
020     031601348X
020     9780316014038
020     0316014036
037     a ǂb ad
043     n-us---
050 0 0 HD9199.U54 ǂb S733 2007
082 0 0 338/.17373 ǂ2 22
084     15.87 ǂ2 bcl
084     7,26 ǂ2 ssgn
084     F719.3 ǂ2 clc
090     ǂb
049     OCLC
100 1   Clark, Taylor, ǂd 1979- ǂ1 http://viaf.org/viaf/161599025 ǂ1
http://id.loc.gov/rwo/agents/n2007022182
245 1 0 Starbucked : ǂb a double tall tale of caffeine, commerce, and culture / ǂc Taylor
Clark.
250     First edition.
264   1 New York ǂb Little, Brown ǂc 2007.
264   4 ǂc ©2007
300     297 pages ǂc 25 cm
336     text ǂb txt ǂ2 rdacontent
337     unmediated ǂb n ǂ2 rdamedia
338     volume ǂb nc ǂ2 rdacarrier
504     Includes bibliographical references (pages 275-289) and index.
505 0   Introduction: The experiment -- The rise of the mermaid. Life before lattes ; A
caffeinated craze ; The siren's song ; Leviathan -- Getting steamed. Storm brewing ; A fair
trade? ; What's in your cup ; Green-apron army ; The Seattle colonies -- Epilogue: The last
drop.
520     The first book to explore the rise of the Starbucks Corporation--and the caffeine-
crazy culture that fueled its success--combines investigative heft with witty cultural
observation in telling the story of how the coffeehouse movement changed our everyday
lives, from our evolving neighborhoods and workplaces to the ways we shop, socialize, and
self-medicate. Journalist Clark provides an objective, meticulously reported look at the
volatile issues like gentrification and fair trade that distress activists and coffee
zealots alike. Through a cast of characters that includes coffee-wild hippies, business
sharks, slackers and Hollywood trendsetters, Clark explores how America transformed into a
nation of coffee gourmets in only a few years, how Starbucks manipulates psyches and social
```

De-duplicated record

Where does this leave us?

Duplicate Detection & Resolution

- 1 million duplicates removed

Member Merge Program

- 20,000 duplicates removed

OCLC Quality Staff Activities

- 12,000 duplicates removed

Over **1 million duplicate** records removed between July 2022 and March 2023.

Source: OCLC, MQ Accumulated Stats | Data as of March 2023.

Where does this leave us?

Duplicate Detection & Resolution

- 1 million duplicates removed

Member Merge Program

- 20,000 duplicates removed

OCLC Quality Staff Activities

- 12,000 duplicates removed

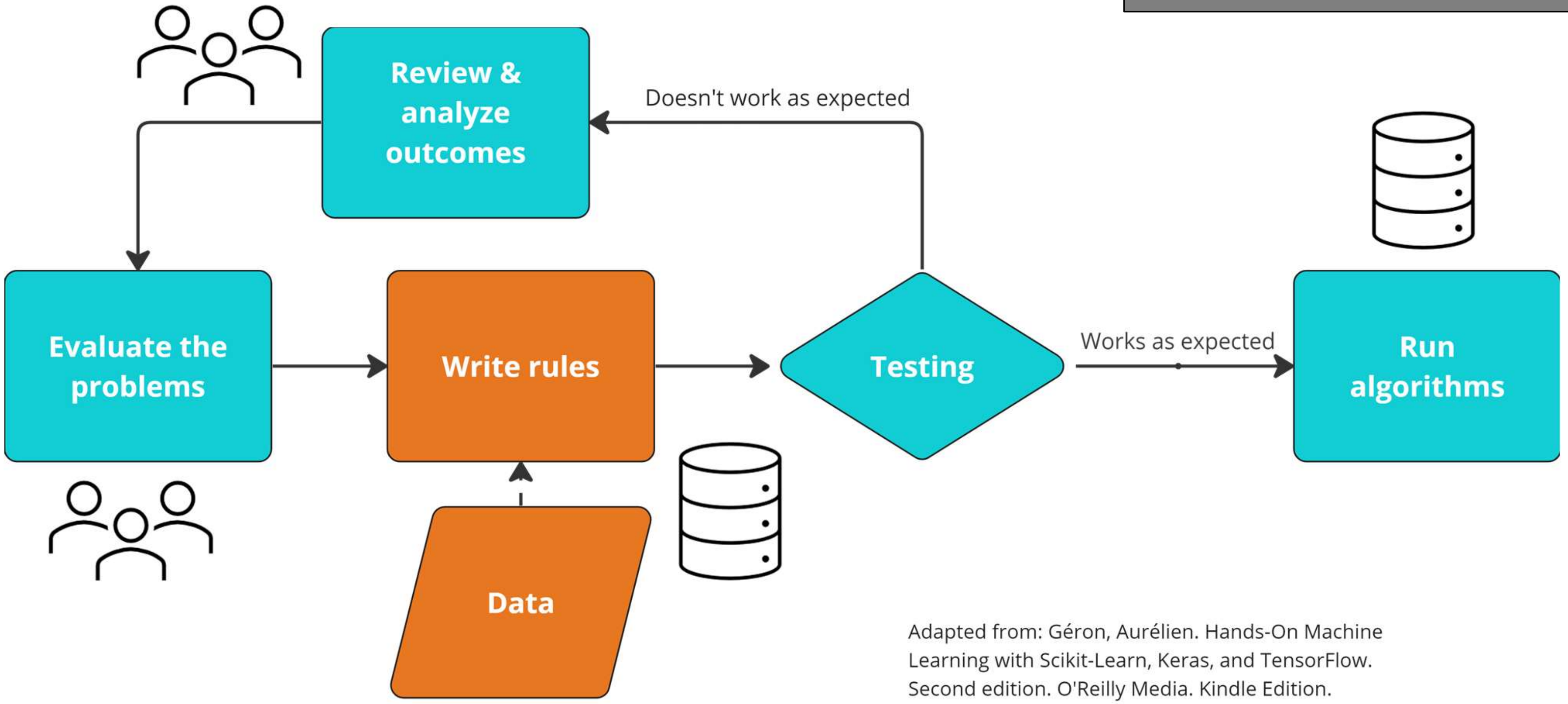
Over **1 million duplicate** records removed between July 2022 and March 2023.

…but we all know there are more out there.

Source: OCLC, MQ Accumulated Stats | Data as of March 2023.

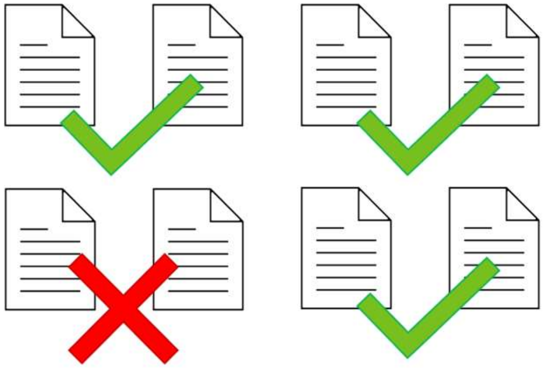
Machine Learning

The science (and art) of programming computers to learn from data



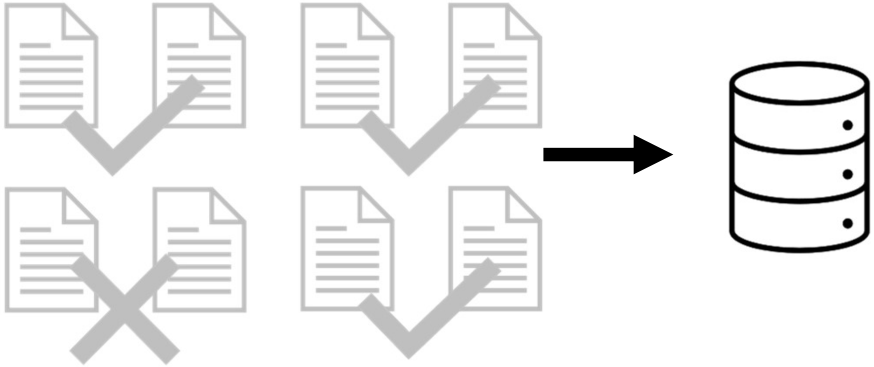
Adapted from: Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. Second edition. O'Reilly Media. Kindle Edition.

Machine Learning



Training set

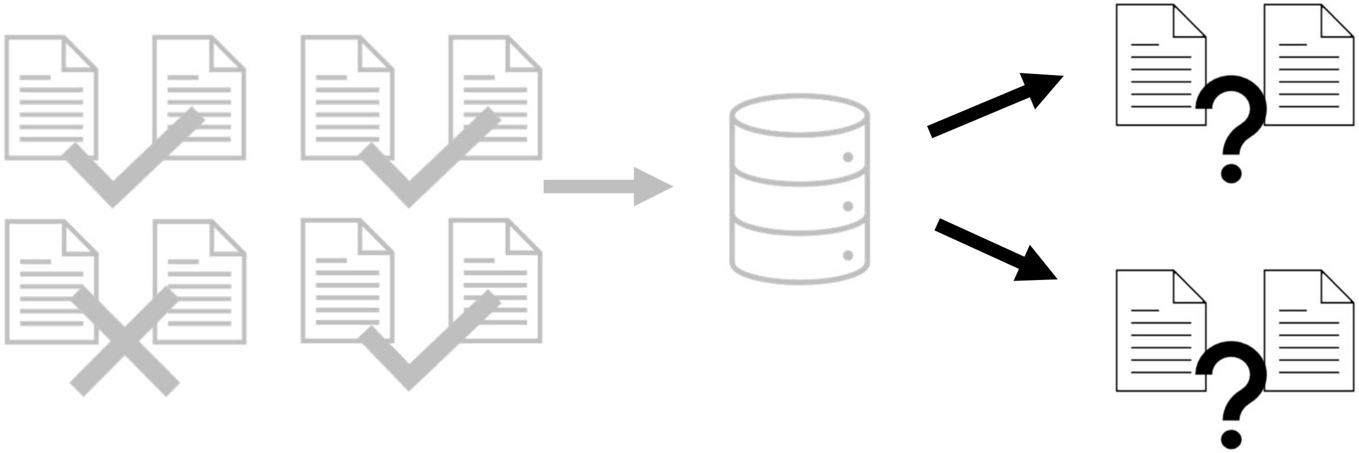
Machine Learning



Training set

“Learns” differences

Machine Learning

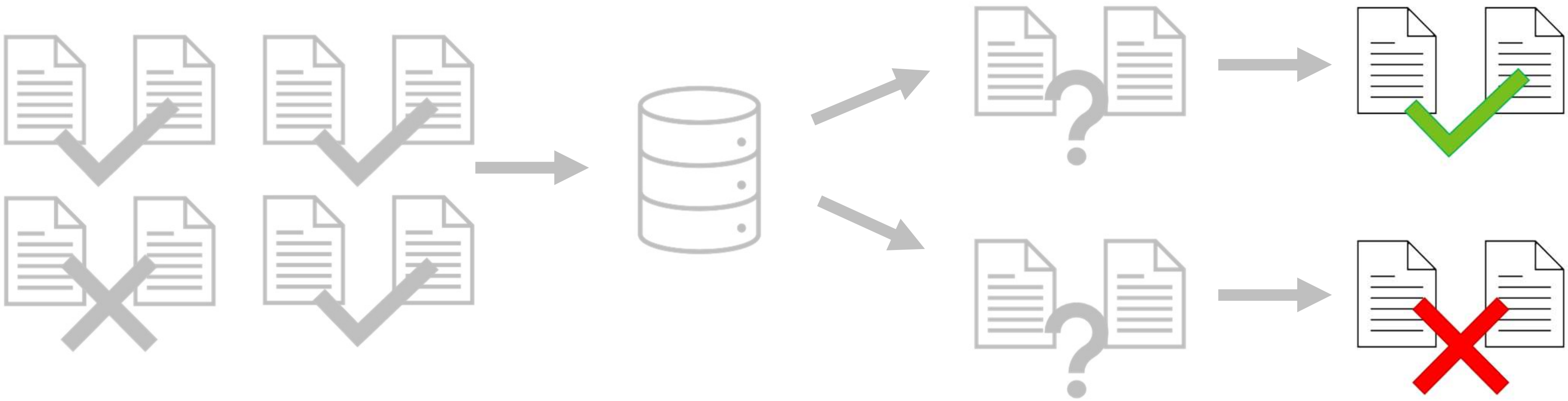


Training set

“Learns” differences

New data

Machine Learning



Training set

"Learns" differences

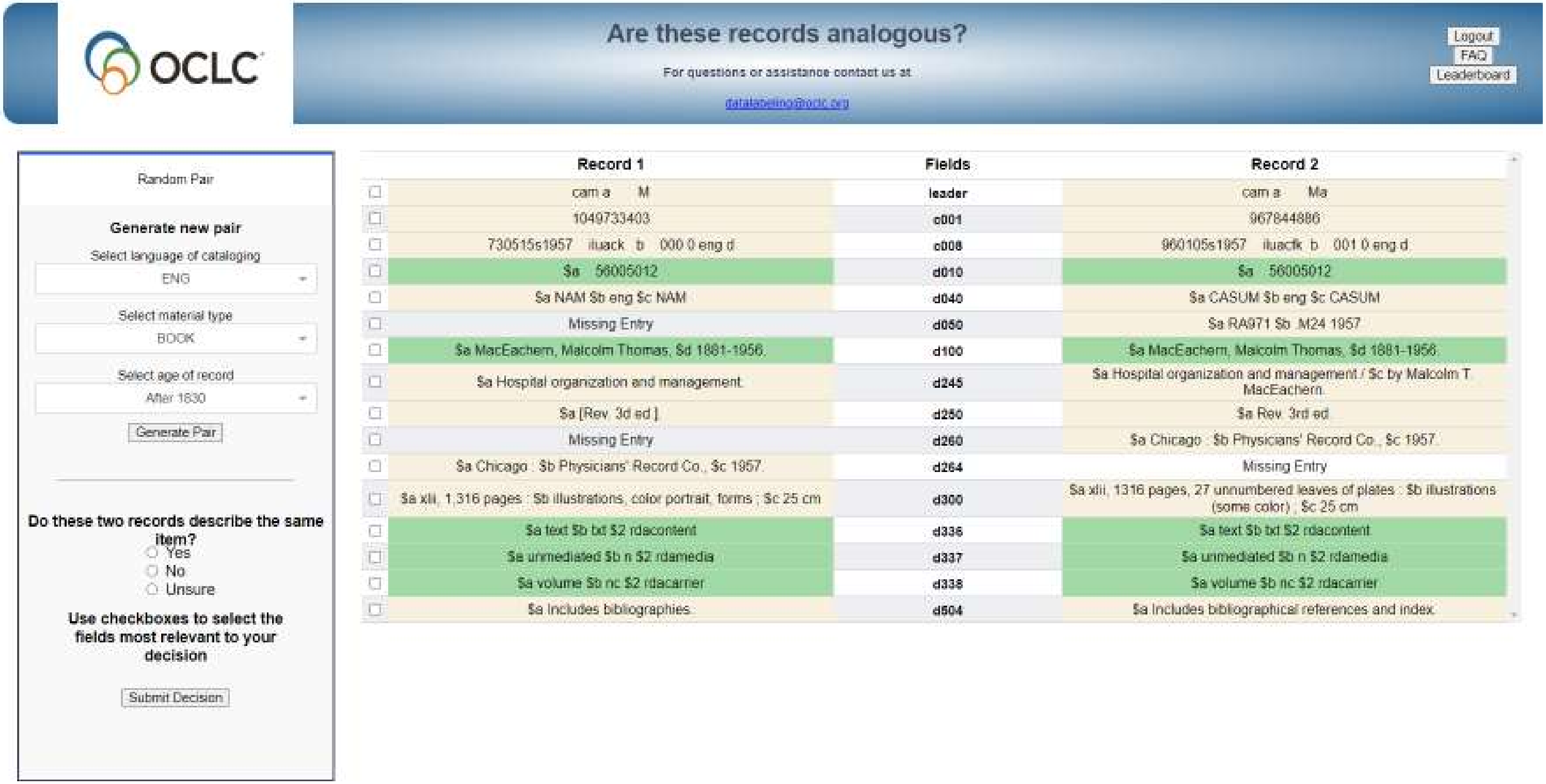
New data

Results

Machine Learning projects

- OCLC Data Labelling Project

  Objective: *Tap into the expertise of the cataloging community to validate and enhance our model's understanding of duplicate records, which ultimately improves the quality of WorldCat for the entire cooperative and library community*
- A sample of 60,000 sets for the tool was generated using MinHash
- The labelled data should be representative of WorldCat in terms of language of item, language of cataloging, material type, etc.
- For more information about MinHash: https://en.wikipedia.org/wiki/MinHash



Are these records analogous?

For questions or assistance contact us at

datalabeling@oclc.org

Random Pair

Generate new pair

Select language of cataloging: ENG

Select material type: BOOK

Select age of record: After 1830

Do these two records describe the same item?

○ Yes
○ No
○ Unsure

Use checkboxes to select the fields most relevant to your decision

	Record 1	Fields	Record 2
□	cam a M	leader	cam a Ma
□	1049733403	c001	967844886
□	730515s1957 iluack b 000 0 eng d	c008	960105s1957 iluacfk b 001 0 eng d
□	$a 58005012	d010	$a 58005012
□	$a NAM $b eng $c NAM	d040	$a CASUM $b eng $c CASUM
□	Missing Entry	d050	$a RA971 $b .M24 1957
□	$a MacEachern, Malcolm Thomas, $d 1881-1956.	d100	$a MacEachern, Malcolm Thomas, $d 1881-1956.
□	$a Hospital organization and management.	d245	$a Hospital organization and management / $c by Malcolm T. MacEachern.
□	$a [Rev. 3d ed.]	d250	$a Rev. 3rd ed.
□	Missing Entry	d260	$a Chicago : $b Physicians' Record Co., $c 1957.
□	$a Chicago : $b Physicians' Record Co., $c 1957.	d264	Missing Entry
□	$a xiii, 1,316 pages : $b illustrations, color portrait, forms ; $c 25 cm	d300	$a xiii, 1316 pages, 27 unnumbered leaves of plates : $b illustrations (some color) ; $c 25 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 Includes bibliographies.	d504	$a Includes bibliographical references and index.

Labelling project stats

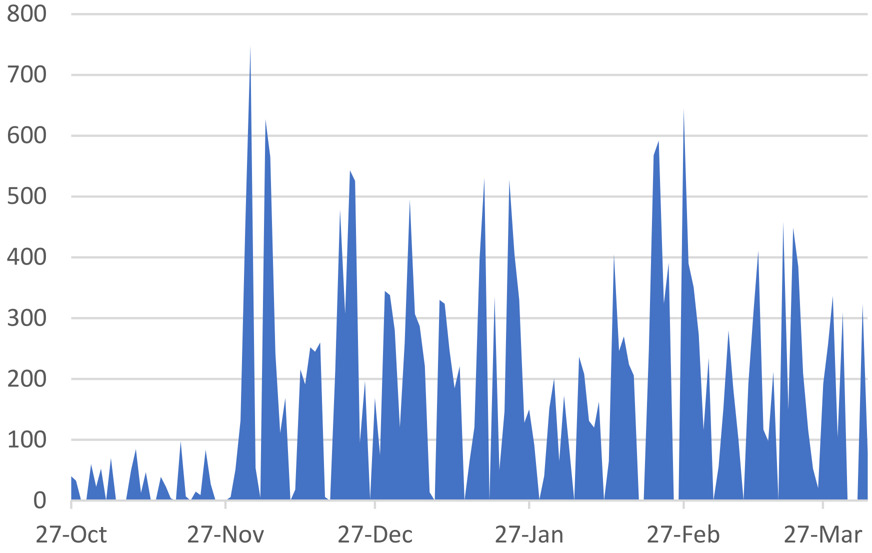
Comparison outcomes

- 40% → the pair IS a duplicate
- 60% → the pair IS NOT a duplicate

Top 5 fields used for decisions

- 245 – Title statement
- 300 – Physical description
- 260 – Publication, etc. statement
- 100 – Main entry - personal name
- 264 – Publication, etc. statement

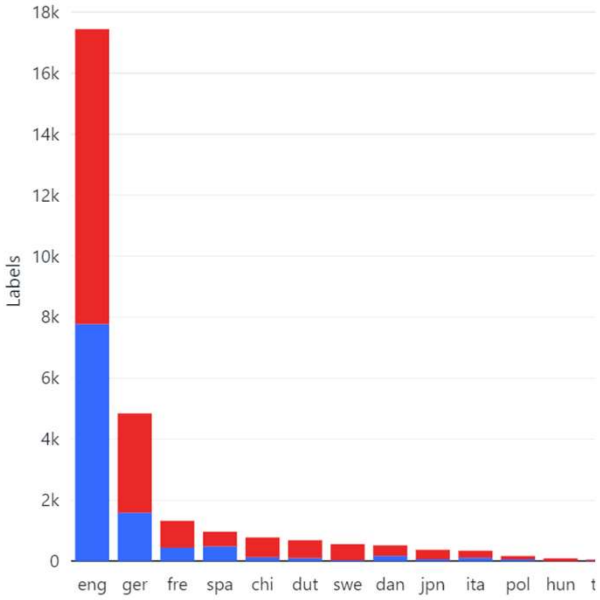
Daily contribution counts



Source: OCLC, Data Science| Data as of April 2023.

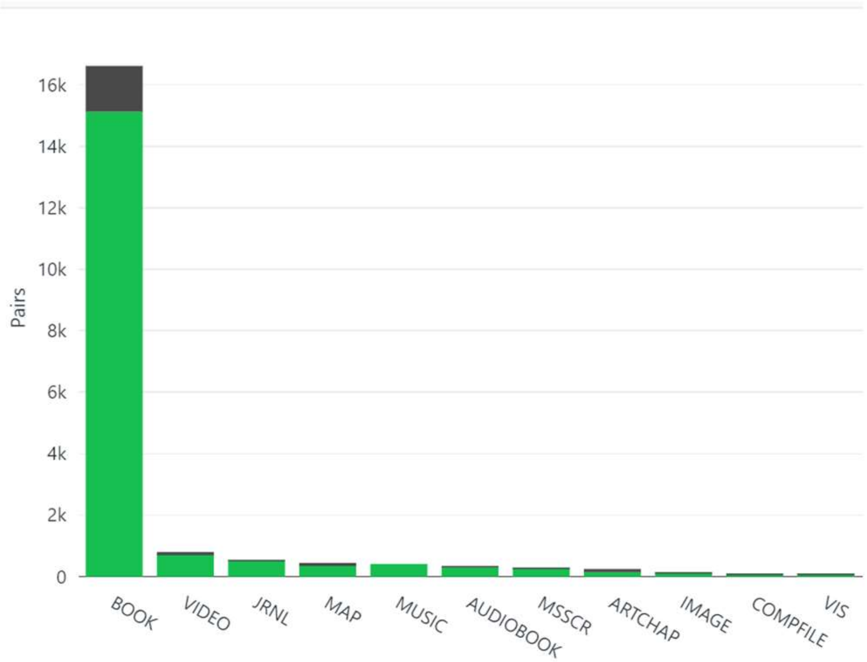
Labelling project stats

Language of cataloging



Source: OCLC, Data Science| Data as of April 2023.

Item format



Progress so far

- Group 1—Exact duplicates
  - Primary match keys are normalized and compared
- Group 2—Vendor duplicates
  - Records from identified vendors with looser criteria
- Group 3—Model duplicates
  - Labelled data used to train model

Considerations

- Like rule-based instructions, ML isn’t a catch-all solution
  - Models provide a degree of accuracy
- Accepting the degree of uncertainty
  - Is cleaning up millions of duplicate records worth a thousand or so incorrect merges?
- Shift in thinking
  - A rule for every situation is not possible, especially complex data sets
  - Use ML to fill in gaps
  - Experts still involved in training set and evaluation

On the call today

Hayley Moreno

Senior Data Analyst

Laura Ramsey

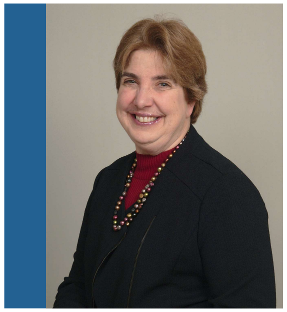
Senior Metadata Operations Manager

Nathan Putnam

Director, Data Quality & Governance

Jenny Toves

Senior Technical Manager



Becky Dean

Lead Data Analyst

Robin Six

Senior Data Analyst

Thank you!

May Virtual AskQC Office Hours
What's in a name? Descriptive access points overview

Tuesday, 9 May at 10:00 AM Eastern
Thursday, 18 May at 4:00 PM Eastern

Registration and session links available at oc.lc/askqc

Send cataloging policy questions at any time to askqc@oclc.org

Photo by Eric Rothermel on Unsplash