Wikidata and the Penn Deep Backfile

Presenters:

John Mark Ockerbloom

Beth Picknally Camden

Jim Hahn

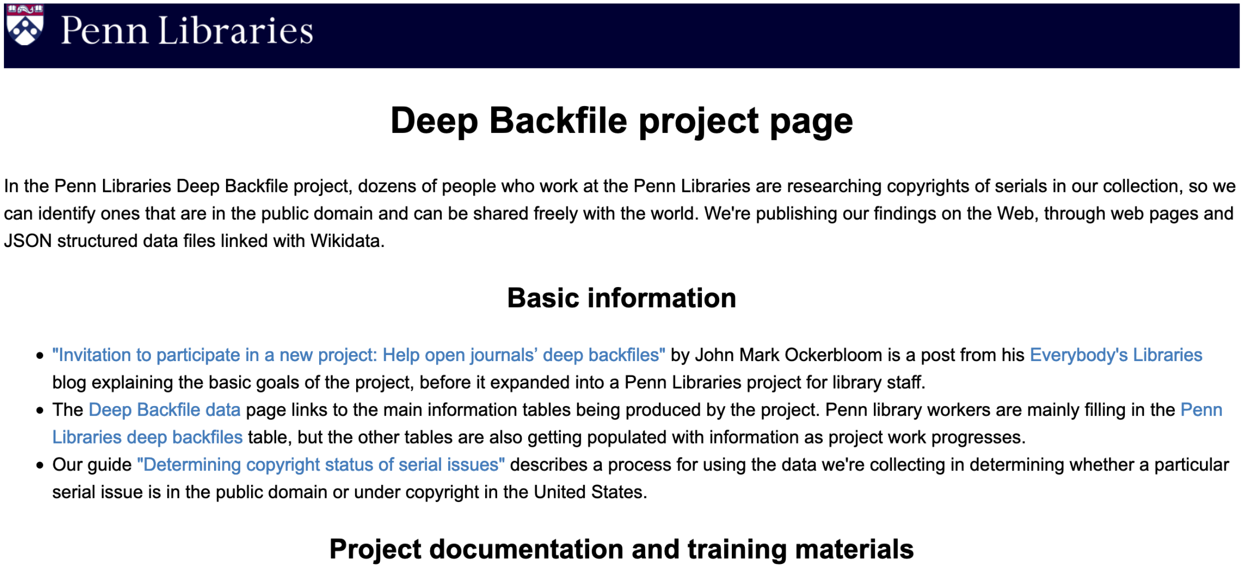
Penn Libraries
University of Pennsylvania

Background

Penn Libraries
University *of* Pennsylvania

Deep backfile project

Workers throughout Penn Libraries research copyright and free online availability of serials we hold, so they can be shared with the world.



Deep Backfile project page

In the Penn Libraries Deep Backfile project, dozens of people who work at the Penn Libraries are researching copyrights of serials in our collection, so we can identify ones that are in the public domain and can be shared freely with the world. We're publishing our findings on the Web, through web pages and JSON structured data files linked with Wikidata.

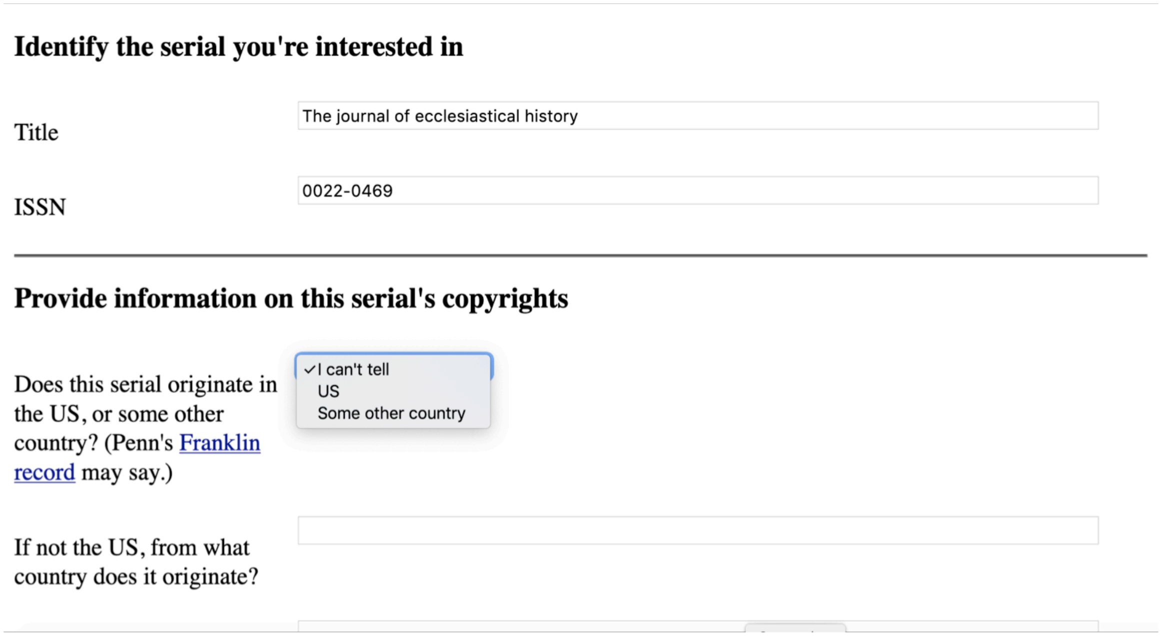
Basic information

- "Invitation to participate in a new project: Help open journals' deep backfiles" by John Mark Ockerbloom is a post from his Everybody's Libraries blog explaining the basic goals of the project, before it expanded into a Penn Libraries project for library staff.
- The Deep Backfile data page links to the main information tables being produced by the project. Penn library workers are mainly filling in the Penn Libraries deep backfiles table, but the other tables are also getting populated with information as project work progresses.
- Our guide "Determining copyright status of serial issues" describes a process for using the data we're collecting in determining whether a particular serial issue is in the public domain or under copyright in the United States.

Project documentation and training materials

https://onlinebooks.library.upenn.edu/cce/db/

A variety of roles for a variety of workers



Identify the serial you're interested in

Title: The journal of ecclesiastical history

ISSN: 0022-0469

Provide information on this serial's copyrights

Does this serial originate in the US, or some other country? (Penn's [Franklin record] may say.)

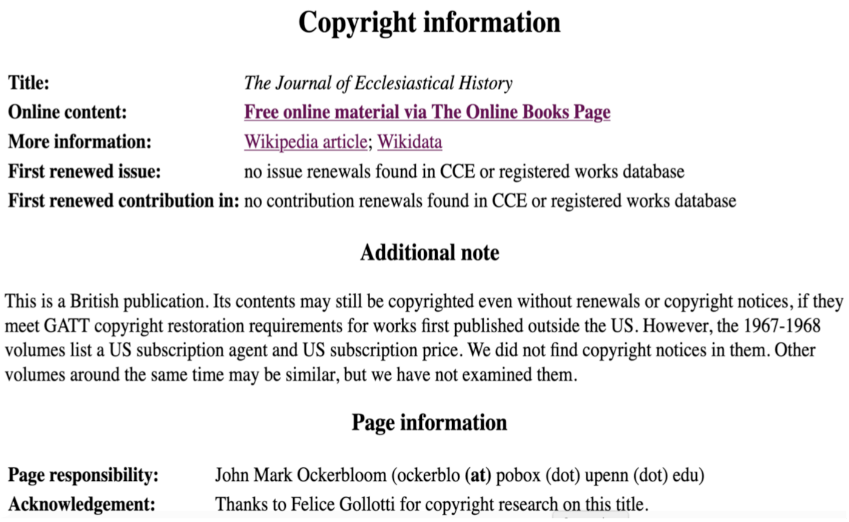
✓ I can't tell
US
Some other country

If not the US, from what country does it originate?

- We trained workers to search for copyright information in online databases and resources and fill out questionnaires on what they find.

- Work can be done from home, and in discrete chunks of time.

After the questionnaire, more specialized work



Copyright information

Title:	*The Journal of Ecclesiastical History*
Online content:	**Free online material via The Online Books Page**
More information:	Wikipedia article; Wikidata
First renewed issue:	no issue renewals found in CCE or registered works database
First renewed contribution in:	no contribution renewals found in CCE or registered works database

Additional note

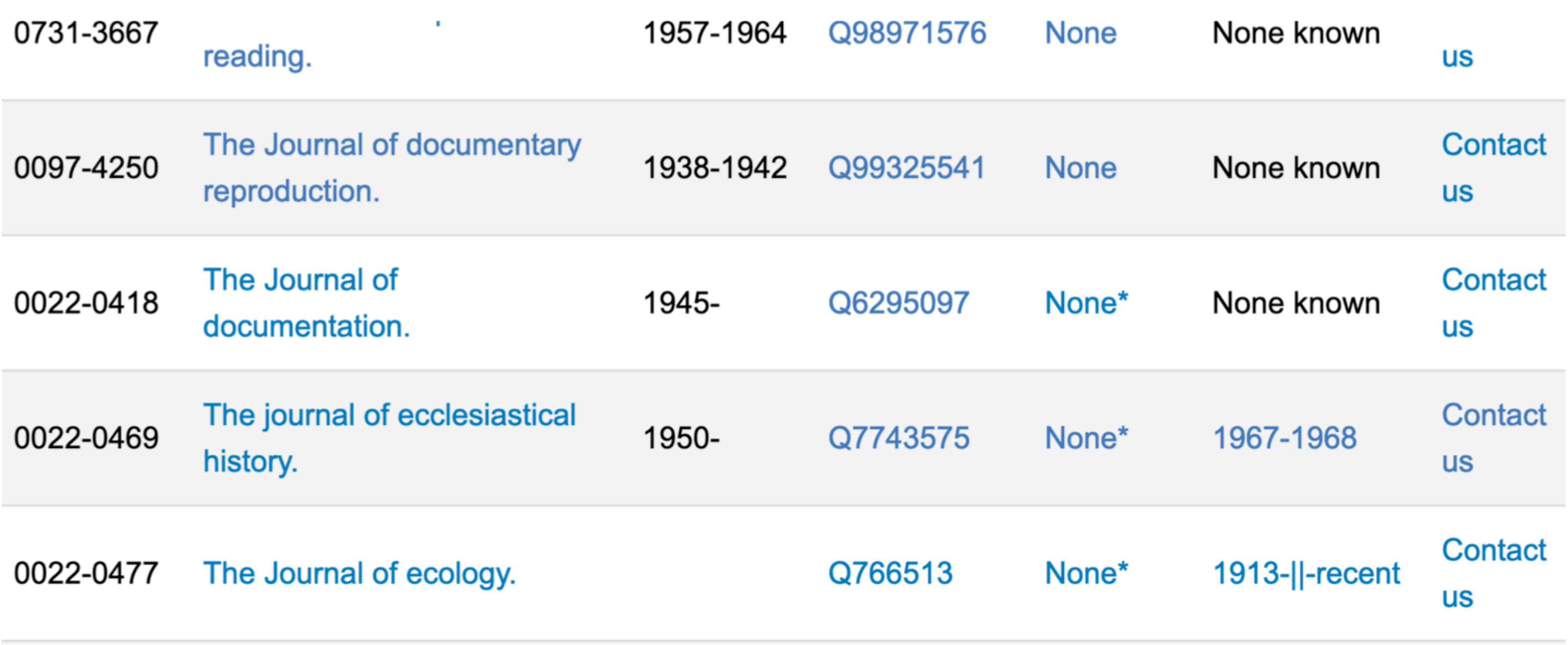
This is a British publication. Its contents may still be copyrighted even without renewals or copyright notices, if they meet GATT copyright restoration requirements for works first published outside the US. However, the 1967-1968 volumes list a US subscription agent and US subscription price. We did not find copyright notices in them. Other volumes around the same time may be similar, but we have not examined them.

Page information

Page responsibility:	John Mark Ockerbloom (ockerblo **(at)** pobox (dot) upenn (dot) edu)
Acknowledgement:	Thanks to Felice Gollotti for copyright research on this title.

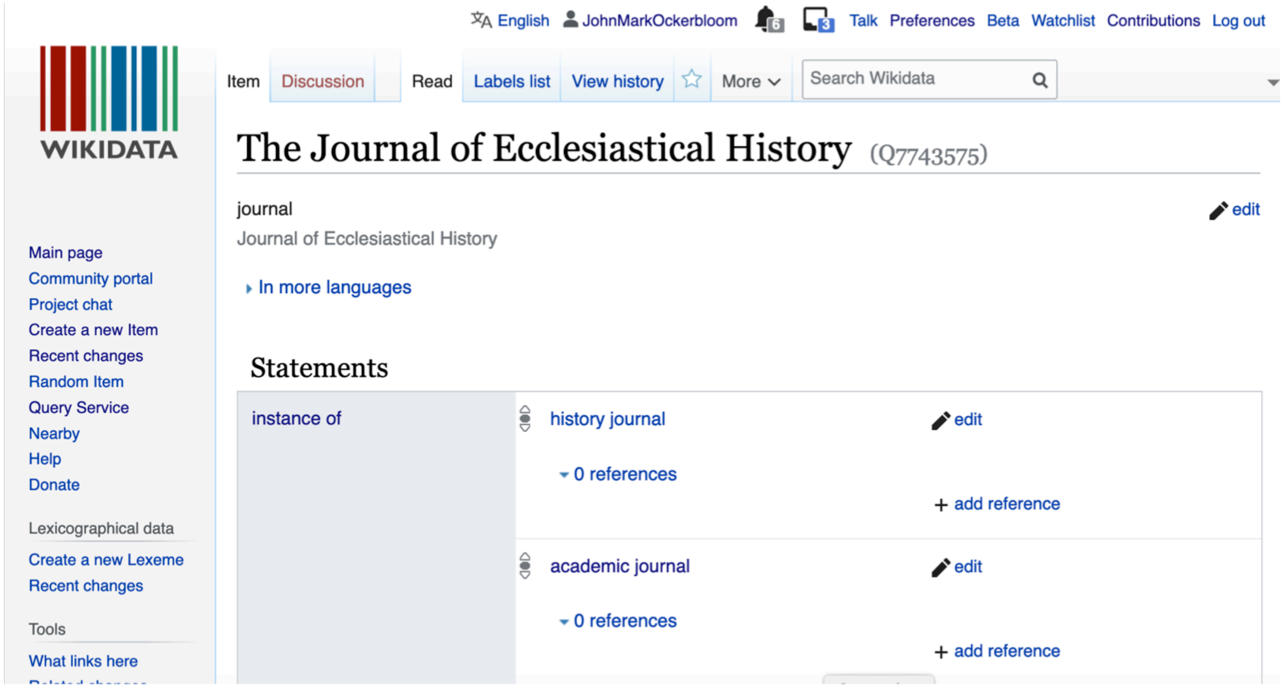
- Verify questionnaire answers
- Create copyright information record (JSON file generating a web page)
- Create Online Books listing, if free issues are online
- Link to Wikidata, either by
  - Adding identifiers to existing Wikidata record for the serial
  - Or, creating a Wikidata record for the serial and adding IDs to it

Wikidata as a hub for identifiers and information



0731-3667	reading.	1957-1964	Q98971576	None	None known	us
0097-4250	The Journal of documentary reproduction.	1938-1942	Q99325541	None	None known	Contact us
0022-0418	The Journal of documentation.	1945-	Q6295097	None*	None known	Contact us
0022-0469	The journal of ecclesiastical history.	1950-	Q7743575	None*	1967-1968	Contact us
0022-0477	The Journal of ecology.		Q766513	None*	1913-||-recent	Contact us

In the Wikidata record



The Journal of Ecclesiastical History (Q7743575)

journal

Journal of Ecclesiastical History

Statements

instance of	history journal 0 references	edit add reference
	academic journal 0 references	edit add reference

Wikidata Serial Statements

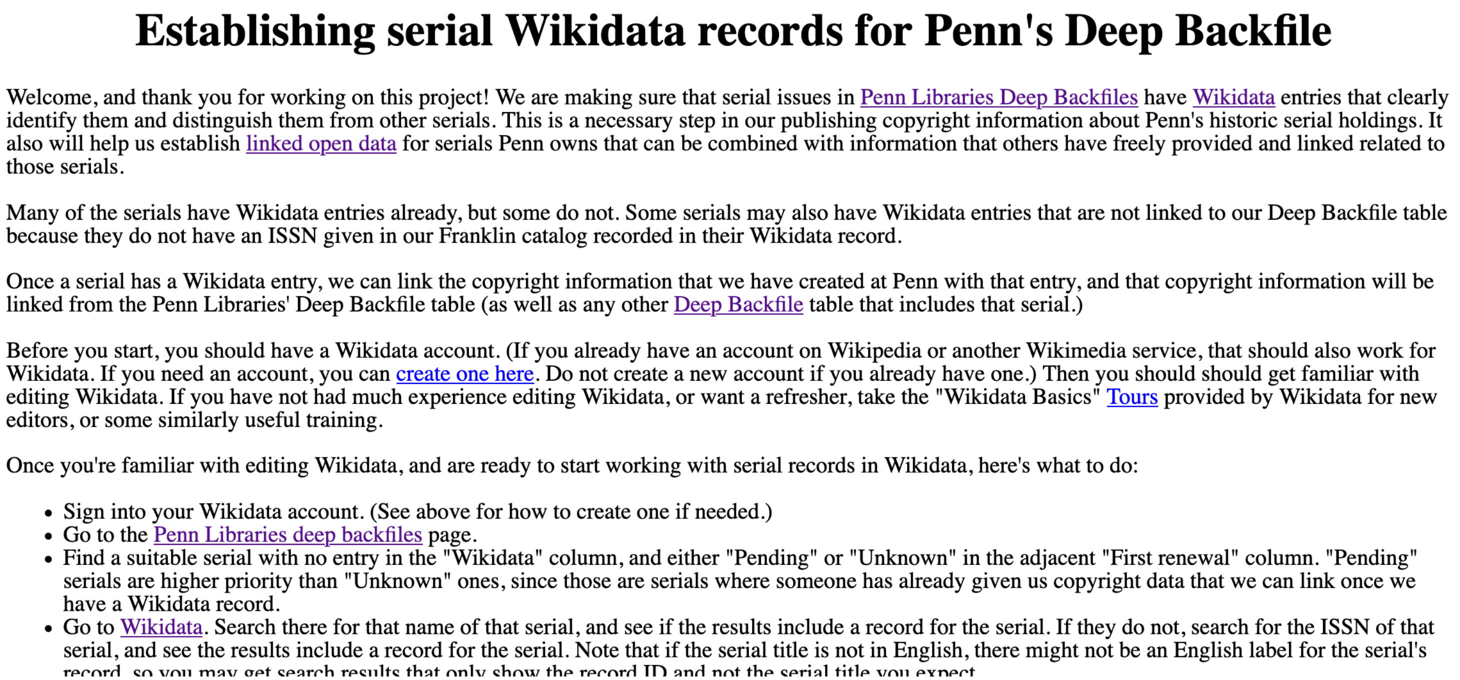
- instance of
- ISSN
- language of work or name
- title
- country of origin
- inception
- dissolved, abolished or demolished date

- replaced by
- replaces

- Online Books Page publication ID

Penn Libraries
University of Pennsylvania

Documentation for editing Wikidata serial records



Establishing serial Wikidata records for Penn's Deep Backfile

Welcome, and thank you for working on this project! We are making sure that serial issues in Penn Libraries Deep Backfiles have Wikidata entries that clearly identify them and distinguish them from other serials. This is a necessary step in our publishing copyright information about Penn's historic serial holdings. It also will help us establish linked open data for serials Penn owns that can be combined with information that others have freely provided and linked related to those serials.

Many of the serials have Wikidata entries already, but some do not. Some serials may also have Wikidata entries that are not linked to our Deep Backfile table because they do not have an ISSN given in our Franklin catalog recorded in their Wikidata record.

Once a serial has a Wikidata entry, we can link the copyright information that we have created at Penn with that entry, and that copyright information will be linked from the Penn Libraries' Deep Backfile table (as well as any other Deep Backfile table that includes that serial.)

Before you start, you should have a Wikidata account. (If you already have an account on Wikipedia or another Wikimedia service, that should also work for Wikidata. If you need an account, you can create one here. Do not create a new account if you already have one.) Then you should should get familiar with editing Wikidata. If you have not had much experience editing Wikidata, or want a refresher, take the "Wikidata Basics" Tours provided by Wikidata for new editors, or some similarly useful training.

Once you're familiar with editing Wikidata, and are ready to start working with serial records in Wikidata, here's what to do:

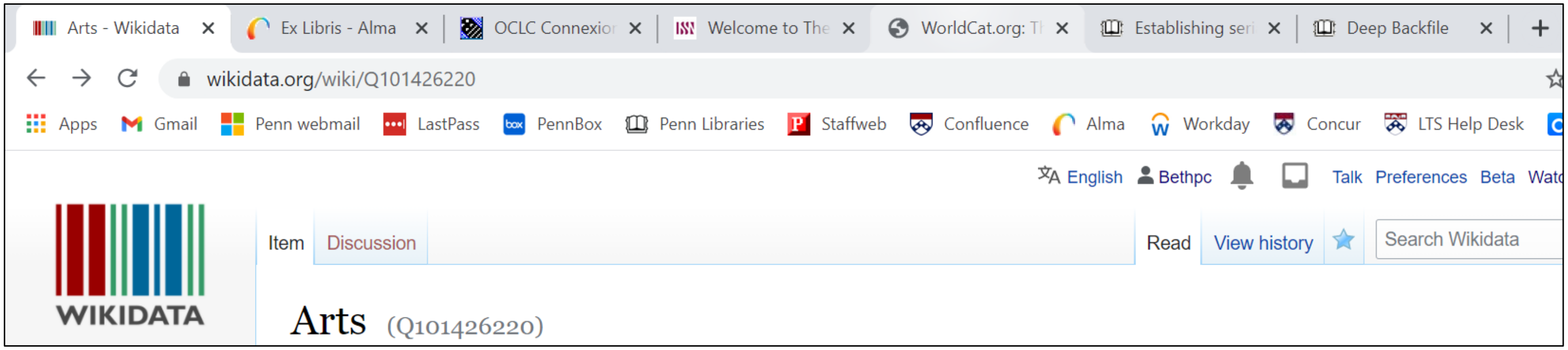
- Sign into your Wikidata account. (See above for how to create one if needed.)
- Go to the Penn Libraries deep backfiles page.
- Find a suitable serial with no entry in the "Wikidata" column, and either "Pending" or "Unknown" in the adjacent "First renewal" column. "Pending" serials are higher priority than "Unknown" ones, since those are serials where someone has already given us copyright data that we can link once we have a Wikidata record.
- Go to Wikidata. Search there for that name of that serial, and see if the results include a record for the serial. If they do not, search for the ISSN of that serial, and see the results include a record for the serial. Note that if the serial title is not in English, there might not be an English label for the serial's record, so you may get search results that only show the record ID and not the serial title you expect.

https://onlinebooks.library.upenn.edu/cce/db/

Challenges

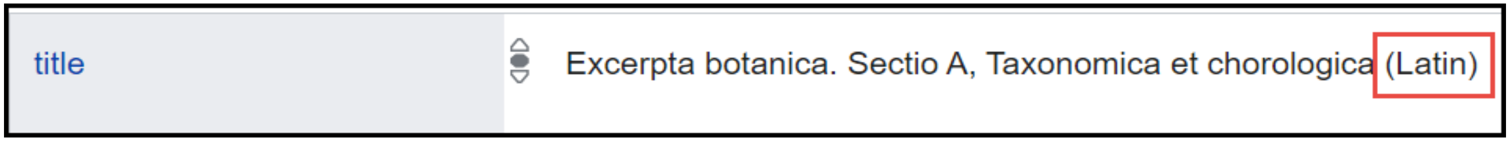
Penn Libraries
UNIVERSITY of PENNSYLVANIA

Lots of tabs


Arts - Wikidata
Ex Libris - Alma
OCLC Connexior
Welcome to The
WorldCat.org: T
Establishing seri
Deep Backfile
wikidata.org/wiki/Q101426220
Apps
Gmail
Penn webmail
LastPass
PennBox
Penn Libraries
Staffweb
Confluence
Alma
Workday
Concur
LTS Help Desk
English
Bethpc
Talk
Preferences
Beta
Watc
WIKIDATA
Item
Discussion
Read
View history
Search Wikidata
Arts (Q101426220)

Languages

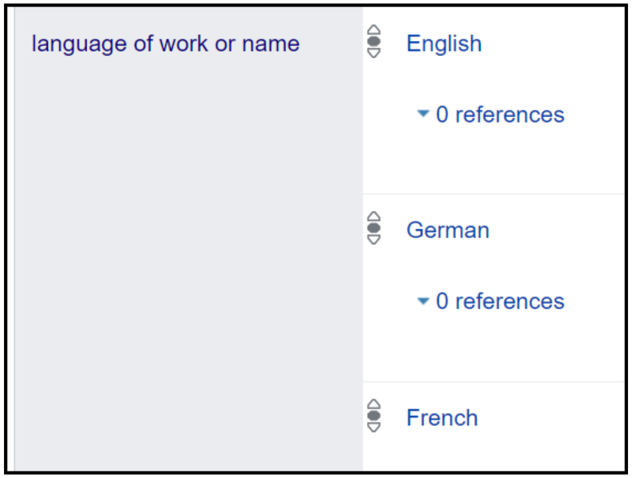
Title in language different than content:



title

Excerpta botanica. Sectio A, Taxonomica et chorologica (Latin)

Content in multiple languages:



language of work or name

English

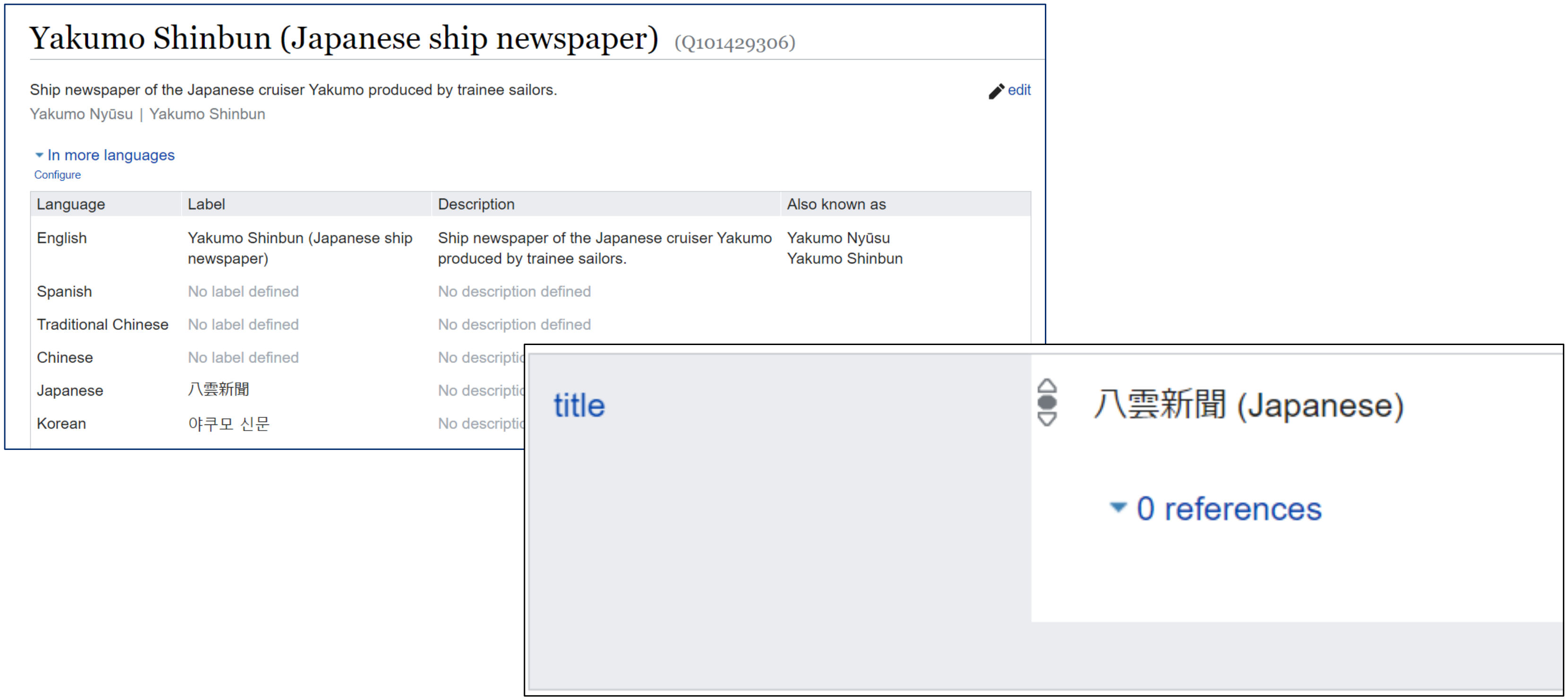
0 references

German

0 references

French

Multiscript entries



Yakumo Shinbun (Japanese ship newspaper) (Q101429306)

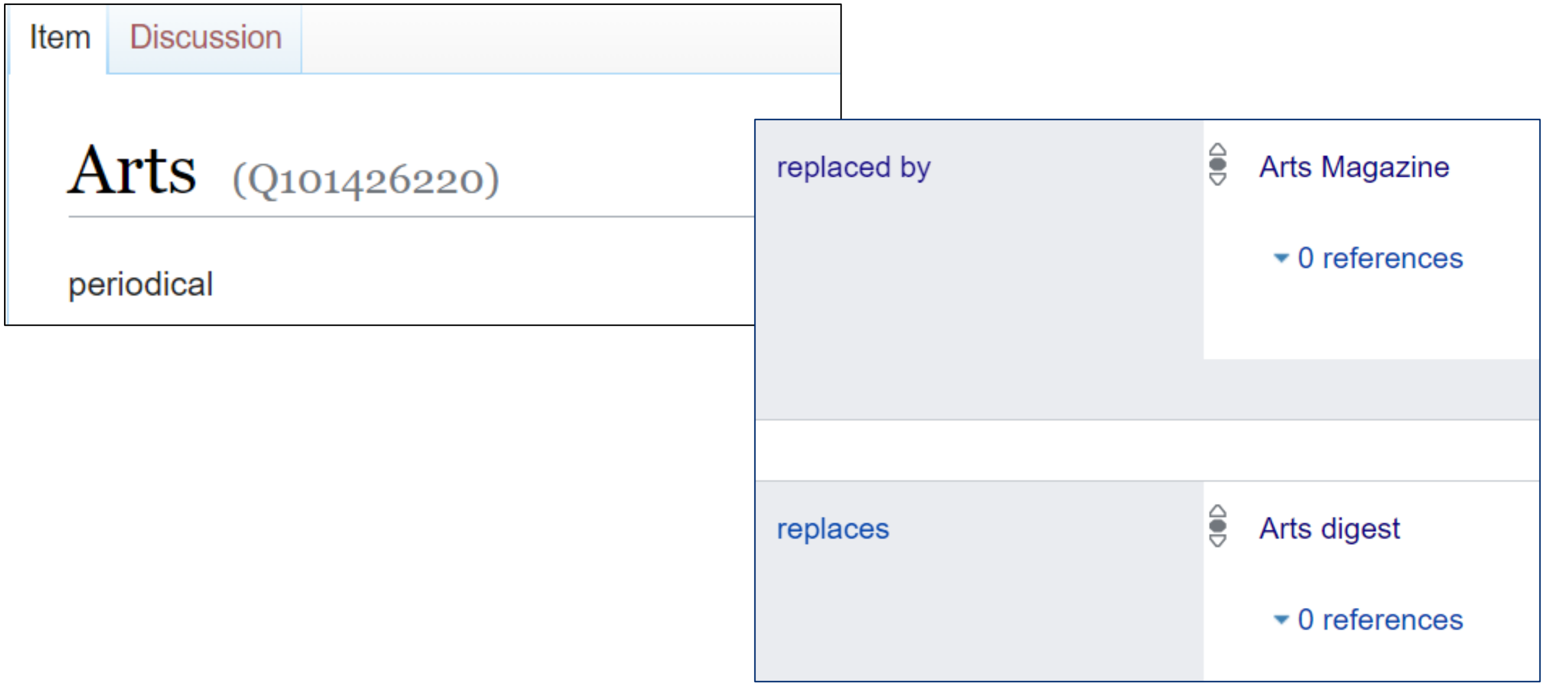
Ship newspaper of the Japanese cruiser Yakumo produced by trainee sailors.

Yakumo Nyūsu | Yakumo Shinbun

Language	Label	Description	Also known as
English	Yakumo Shinbun (Japanese ship newspaper)	Ship newspaper of the Japanese cruiser Yakumo produced by trainee sailors.	Yakumo Nyūsu Yakumo Shinbun
Spanish	No label defined	No description defined	
Traditional Chinese	No label defined	No description defined	
Chinese	No label defined	No descriptio	
Japanese	八雲新聞	No descriptio	
Korean	야쿠모 신문	No descriptio	

title: 八雲新聞 (Japanese)

Title changes



Item | Discussion

Arts (Q101426220)

periodical

replaced by — Arts Magazine

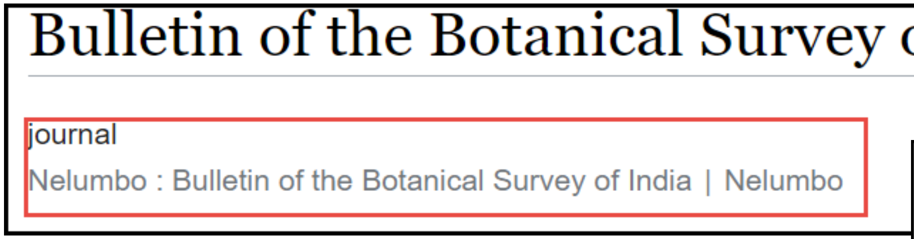
▾ 0 references

replaces — Arts digest

▾ 0 references

Existing Wikidata Entry

Existing Wikidata record lacked our ISSN; one record with multiple titles:

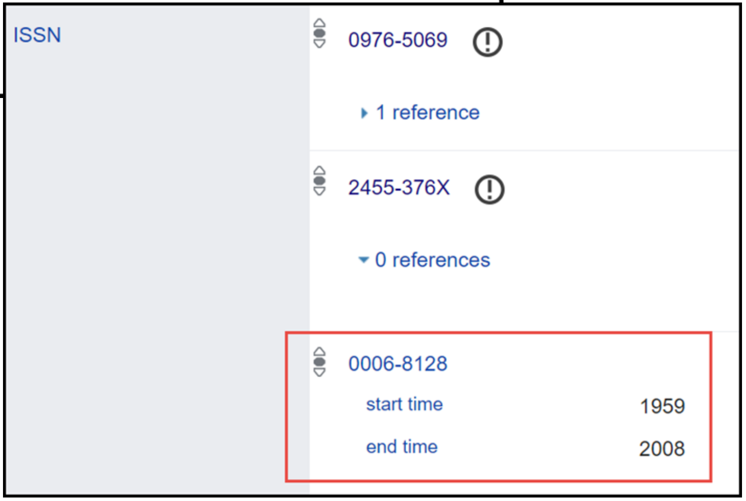


Bulletin of the Botanical Survey of India (Q5735536)

journal

Nelumbo : Bulletin of the Botanical Survey of India | Nelumbo

Added ISSN with date:



ISSN

0976-5069

1 reference

2455-376X

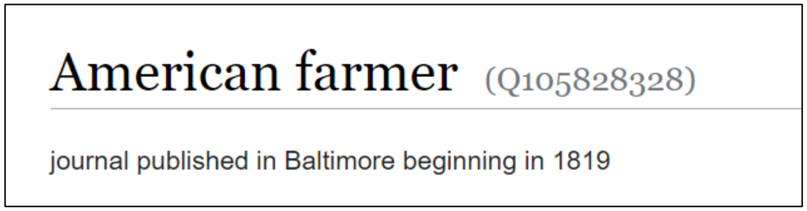
0 references

0006-8128

start time 1959

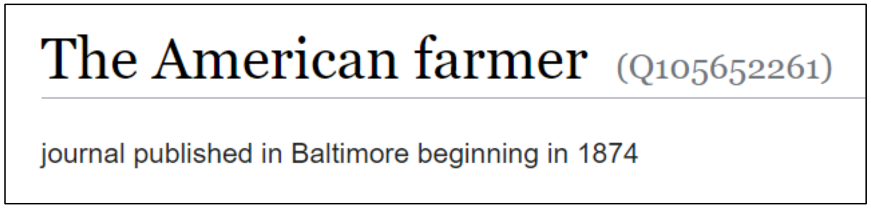
end time 2008

Similar titles



American farmer (Q105828328)

journal published in Baltimore beginning in 1819



The American farmer (Q105652261)

journal published in Baltimore beginning in 1874

Batch Processing

Using OpenRefine

Penn Libraries
University *of* Pennsylvania

OpenRefine Outline

Alma>MarcEdit>OpenRefine>Wikidata

1. MarcEdit Z39.50 retrieve MARC records by ISSN from Penn Alma database. Remove duplicate titles from gathered MARC records.

2. Extract desired fields from records transforming into tab-delimited data using MarcEdit

3. Upload CSV to OpenRefine and reconcile ISSN and title to see if any resources are already in Wikidata.

OpenRefine Outline

4. Clean/organize data: for country codes / language

Cleaning 035: Split into multiple columns; remove columns with irrelevant values; find replace (OCoLC)/null

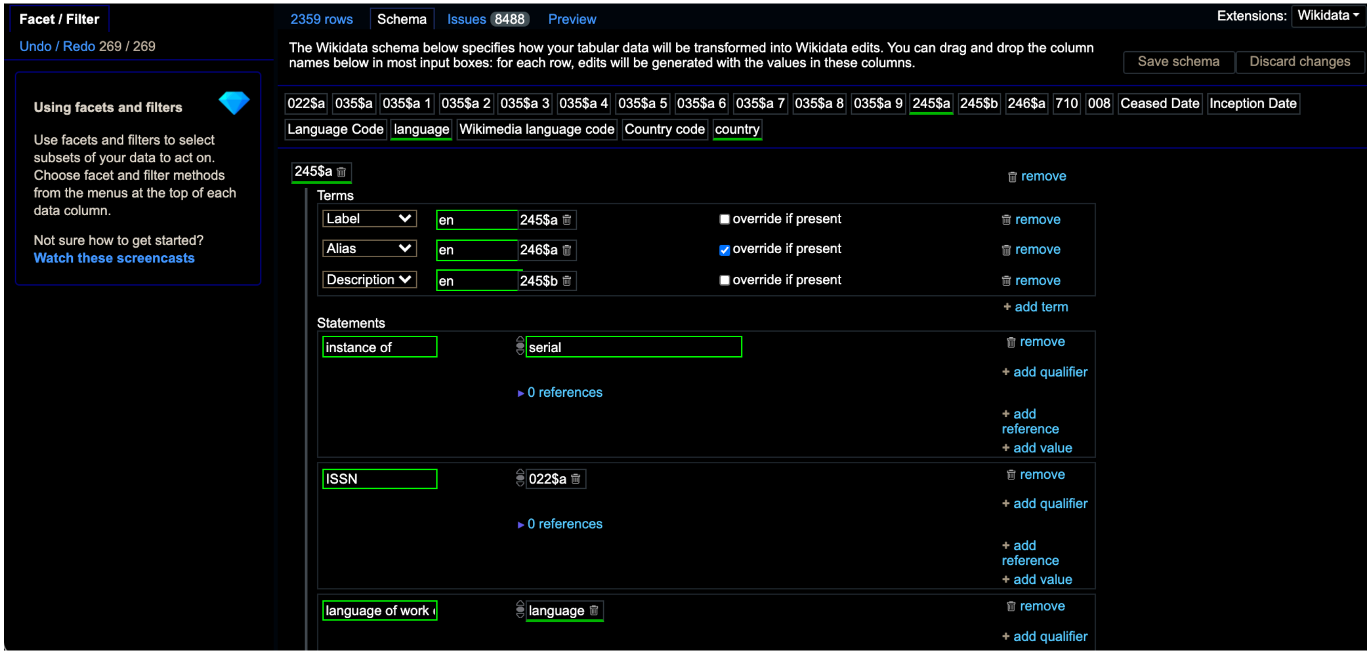
OpenRefine Outline

5. Reconcile Country of Origin, Language

6. Get Wikimedia language code for languages (“Add columns from reconciled values”)

7. Create schema

Defining schema for the serials



Facet / Filter

Undo / Redo 269 / 269

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

2359 rows | Schema | Issues 8488 | Preview

Extensions: Wikidata

The Wikidata schema below specifies how your tabular data will be transformed into Wikidata edits. You can drag and drop the column names below in most input boxes: for each row, edits will be generated with the values in these columns.

Save schema | Discard changes

022$a | 035$a | 035$a 1 | 035$a 2 | 035$a 3 | 035$a 4 | 035$a 5 | 035$a 6 | 035$a 7 | 035$a 8 | 035$a 9 | 245$a | 245$b | 246$a | 710 | 008 | Ceased Date | Inception Date | Language Code | language | Wikimedia language code | Country code | country

245$a — remove

Terms

- Label — en — 245$a — □ override if present — remove
- Alias — en — 246$a — ☑ override if present — remove
- Description — en — 245$b — □ override if present — remove

+ add term

Statements

- instance of — serial — remove — + add qualifier — ▸ 0 references — + add reference — + add value
- ISSN — 022$a — remove — + add qualifier — ▸ 0 references — + add reference — + add value
- language of work (— language — remove — + add qualifier

Generate QuickStatements File

Quickstatements file can be evaluated prior to batch loading...

```
CREATE
LAST    Len    "Admission requirements of American medical colleges"
LAST    P31    Q2217301
LAST    P236    "0271-6526"
LAST    P407    Q1860
LAST    P495    Q30
LAST    P1476    en:"Admission requirements of American medical colleges"
```

For quality assurance:

avoid batch loading any duplicate titles, or any statements that contain the same ISSN. Refer duplicate titles or ISSNs for review. Additional rounds of character parsing in titles may be necessary to remove the AACR2 back slashes in titles and other extraneous non-title punctuation.

Questions?