Center for Research Libraries
GLOBAL RESOURCES NETWORK

JSTOR Catalog Record De-duplication Project

Report to PAN, June 22, 2018

By Stephen Early and Amy Wood

At the June 2017 PAN meeting, CRL reported frustration at the number of duplicate records in OCLC's WorldCat database for titles being retained by print archiving programs. The duplicate records make it difficult to easily know the number of copies held in libraries and to designate an authoritative record. CRL wondered if we could create guidelines and cost out a significant de-duplication project with the ultimate goal of getting the community to support the de-duplication of significant titles—those that were being committed to print retention programs/collections.

CRL decided to use the JSTOR titles as a testbed. JSTOR records were chosen because it was an easy-to-define group of titles and any effort we would make to de-duplicate the records would support CRL's work with its JSTOR print archive. Also, the effort would be the perfect complement to JSTOR's commitment to preserving and managing the print material. JSTOR requests ISSN assignments to all print titles that do not already have them. So, we have digital copies for access, trustworthy print repositories to physically care for the print copies, an internationally recognized authoritative identification number to help us manage the print titles, but no easy way to tell how many copies are still out there and no single shared bibliographic description of the work because we have failed to do the last step: reduce the duplication of catalog records.

Prior to the 2017 PAN meeting, CRL performed an API batch search of OCLC's WorldCat database for JSTOR records using the ISSN numbers listed in JSTOR's list of journal titles found on the JSTOR website. The search targeted the records that had the ISSN in the 022 subfield a. At the time we performed the search there were 3,097 ISSNs for the JSTOR titles. Those ISSNs retrieved over 78,000 records. As we found later, there are significant numbers of JSTOR title catalog records in WorldCat that do not have an ISSN in the 022 field. Less than 1% of the titles had a single record.

Reviewing the retrieved records we found about 2,500 monographic records that contained the ISSNs in the 022 subfield a. These records were divided into two groups: serial records that had been accidentally coded as monographs and articles within the serial issues. Using OCLC's Connexion client and locally developed macros for use in the client, we were able to verify which records were easy to identify as serials; those records were then changed to serial records and updated in WorldCat. The hope was that now coded as serials, the records might be picked up by WorldCat's existing algorithm for de-duplicating records. There were a number of records that needed to be reviewed individually to determine whether they truly cataloged as a serial. If there were records that we could not definitely identify as serials, we left them alone. The catalog records for the articles needed to retain their bib level of monograph, but the ISSN needed to be moved to subfield x in the 773 field. This was also accomplished with a Connexion macro.

Although it was satisfying to be able to make serious progress updating a large number of records, it was a drop in the 78,000-record bucket, and it didn't address the real heart of the problem. The heart of the problem is the

existence of normal duplicate serial records, most of them coded level M or less, which are probably the result of batch loads into WorldCat that failed to find a record match.

To address these records, Stephen Early, CRL's CONSER trained cataloger, decided it would be best to work through a family or group of related titles to get a sense of common and challenging problems, and to create an initial set of guidelines for the work of de-duplication.

The family of related titles began with American College Bulletin, ISSN# 2163-7652, and grew to include nine titles: seven journal titles and two supplements. The two supplements are not listed on JSTOR's list of titles, although they may be included within their parent titles. We did not check for this. They were included in CRL's JSTOR print collection, so the physical copies could be verified and matched with their catalog records. The seven print titles are included within JSTOR's title list as two sets of related titles. JSTOR's title listing may include only linear relationships and this had two separate branches. Screen shots of the JSTOR title groupings and the OCLC ISSN visualization tool representation of this title are included with this report as Attachment 1.

For the nine titles in the American College Bulletin family there were nine print ISSNs and six digital ISSNs. Microform versions use the same ISSNs as the print. For these 15 ISSNs, there were 227 records. There were 101 non-English "language of cataloging" records. Those were ignored. As much as we would like to reduce all duplication, we were hesitant to evaluate non-English cataloging for the best record. All English language-of cataloging records were evaluated by format and the results of de-duplication were:

- Eighty-three duplicate records were reported through OCLC's mechanism for reporting duplicate records.
- Two serial supplement records were kept because of problems that could not be resolved.
- Forty-one records were identified as "preferred". These included records for each format (print, reprint, each microformat, each microformat publisher, and electronic) and "allowable duplicates" (pre-AACR records "latest entry" serial records coded as "S/L 1" in MARC fixed field which are allowed to co-exist with AACR2 and RDA "successive entry" records).

In addition to the duplicate reporting, forty-eight records were enhanced to

- add or correct 022 fields where necessary;
- add $x ISSN to 77x and/or 78x linking fields;
- add 040 "$e pn" to preferred provider neutral electronic records.

No additional enhancements were made (except for any obvious cataloging errors)

Despite being unable to realize visions of one title/one record, we were able to whittle down our family of nine titles from 128 English language records to 43, of which ten were for print (nine PCC, and one pre-AACR latest entry PCC).

A full accounting of records for each ISSN, along with some explanation of initial searching is included as Attachment 2. For next steps, CRL intends to follow the guidelines set by Mr. Early, and catalog enough title families to estimate the time needed, level of experience required and cost of completing the de-duplication of JSTOR records. With a full understanding of cost and time, we will attempt to recruit partners in the effort.

Attachment 1

Screen shots from JSTOR website and OCLC ISSN Visualization Tool to show publication history.

Image copied from JSTOR's web page url: https://www.jstor.org/journal/amercollbull

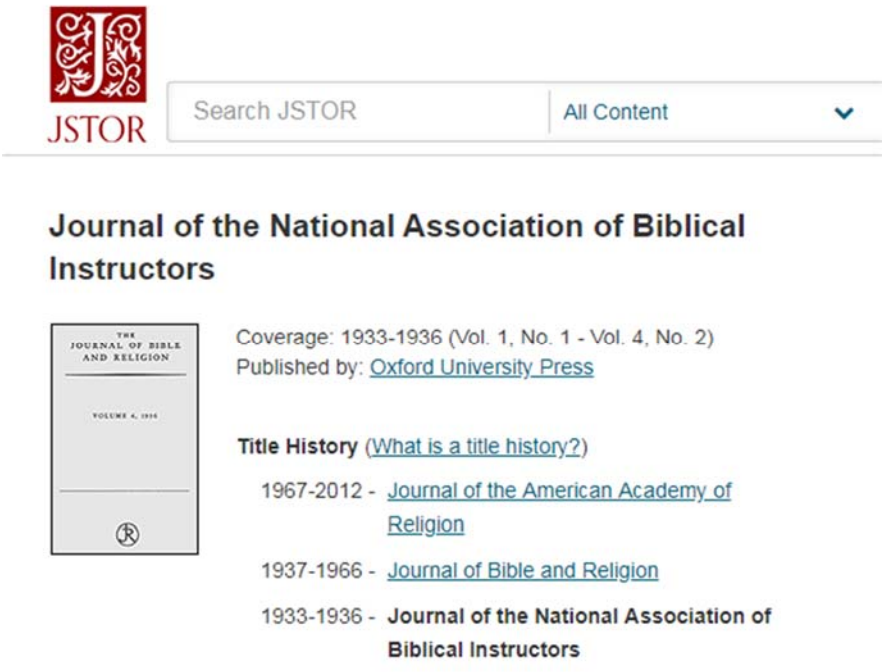

Image copied from JSTOR's web page url: https://www.jstor.org/journal/jnatiassobiblins

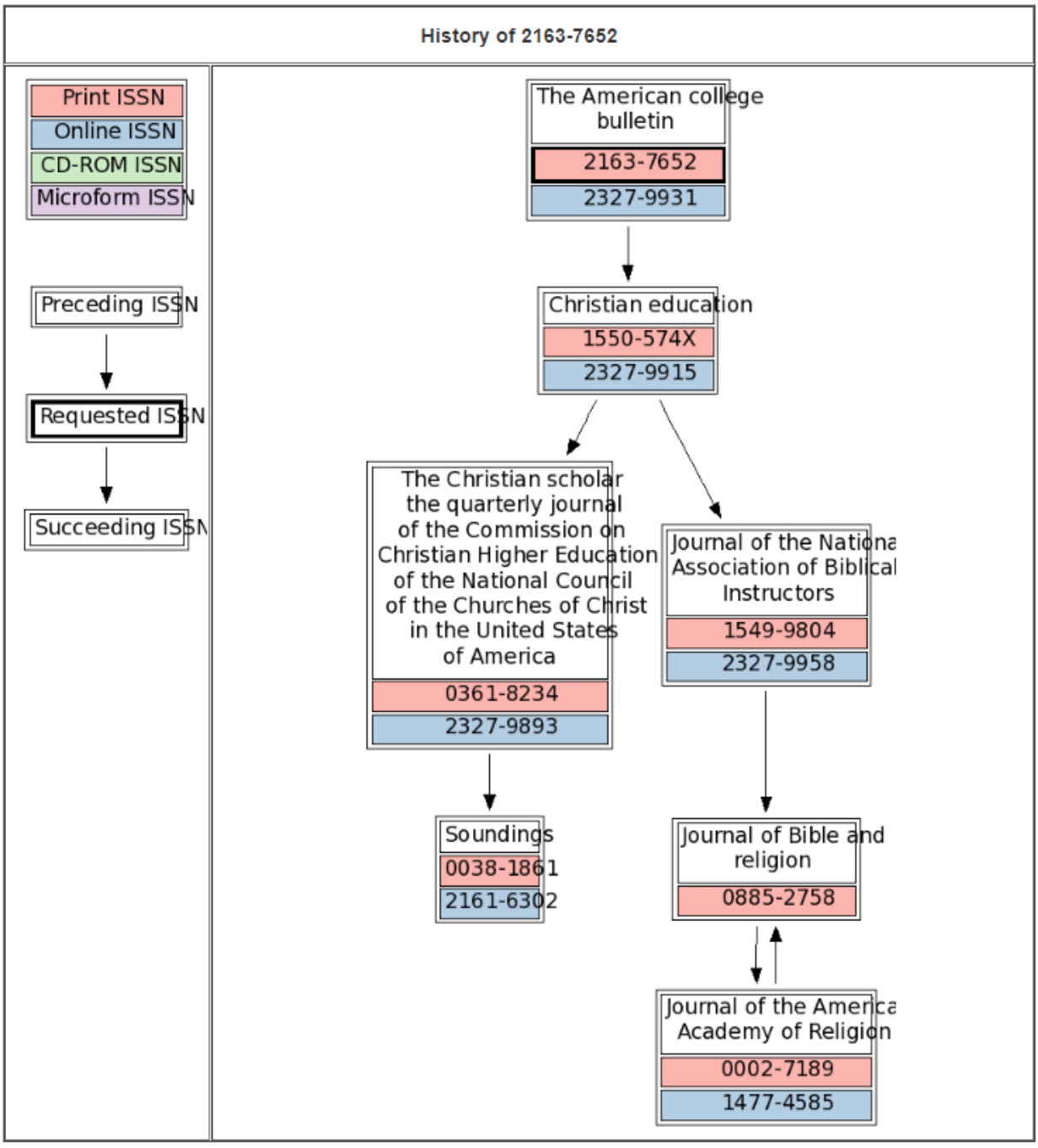

Image copied from results of query by ISSN in OCLC's ISSN visualization tool.

http://worldcat.org/xissn/titlehistory?issn=2163-7652

Attachment 2

ISSN-JSTOR De-duping Research Project
Test case:
Journal of the American College Bulletin (2163-7652) and related titles

Purpose: identify preferred OCLC 040 $b eng JSTOR print, microform, electronic records, or other (mostly reprint). Report all non-preferred records as duplicates.
Non-eng records searched and saved but not otherwise examined.

Titles
9 related titles, 6 titles in JSTOR, 2 titles (supplements) not in JSTOR, 1 title with JSTOR CRL record, but no holdings (not searched in titles.xlsx – see below)

Record sources
- OCLC: Source of actual records for processing.

Fields searched and used for sorting
- Search criteria:
  - Initial searching of both titles.xlsx and OCLC was ISSN only, which captures 022 $a and $y. $l not captured by ISSN search. ISSN-L is a separate search.
  - 156 records captured from ISSN search. ISSN-L search captured 21 additional relevant eng records, of which 3 are pcc, and therefore essential to the project.
  - At end of project, searched for relevant titles lacking ISSN via title. 35 additional relevant eng records found. 21 additional non-eng records found.
- Identification within single ISSN
  - ISSN
  - Format
  - Cataloging level (ELvl)
  - Language
  - Aacr? (pre-AACR2 latest entry records)
  - Filmer: entered manually
- Macro
  - Macro searched each record, starting with the first open one, for above information and added the above as codes in MyStatus field, which can handle 40 characters max.
  - Proved invaluable for record sorting.
- Identification within family of titles
  - This was trickier. Title sequence of JAAR not linear. Two predecessor title sequences prior to 0002-7189. Chose “Rel” (for “Religion”) as root code with small letters as prefix to distinguish the predecessors and double digit numbers as suffix to reflect linear progression within those sequences. Needed to figure out the full title sequence before devising the code.
- Macro
  - Also tricky. This macro produced a message box where the inputter could manually enter the “Rel” code. Could only properly be run after first macro was run on a specific ISSN.
  - Example: 1st macro run on 51 records with ISSN 0002-7189. As a result, #1479270 was assigned the following to MyStatus:
    - “0002-7189 print pcc eng.”

I then listed the same 51 records and ran the 2nd macro, which produced a message box where I entered pre-selected "Rel" code" "bRel 05." Result for #1479270:
"bRel 05 0002-7189 print pcc eng"

ISSN criteria:

1 ISSN and ISSN-L per print title with ISSN-L almost always same as print ISSN. Same print ISSN may be used for microfilm, microfiche, and print reprint records. Different ISSN for electronic records with ISSN-L matching print ISSN-L

Criteria for preferred records:

Print: PCC records in all cases. One preferred S/L 0 and one preferred S/L 1 record (if present).
Film: One record per micropublisher. PCC or next fullest level. If still uncertain, prefer earlier record – generally followed, but some judgment calls. One preferred S/L 0 and one preferred S/L 1 record (if present).
Fiche: same as film
Electronic: PCC or next highest level if no pcc. One preferred S/L 0 and one preferred S/L 1 record (if present). All preferred records updated to include "$e pn" (provider neutral) in 040 if not already present

Editing criteria for preferred non-aacr records:
Make sure all links to other preferred records include ISSNs (usually $x)

Editing criteria for preferred aacr records:
Make sure all 247s and links to other preferred records include ISSNs (usually $x)

List of JSTOR related titles searched (includes code, III b number, and number of title.xlsx found).

aRel 01
2163-7652
American college bulletin
JSTOR b28189759
0 Titles.xlsx

aRel 02
1550-574X
Christian education (Chicago, Ill.)
JSTOR b28190786
7 Titles.xlsx

aRel 03
0361-8234
Christian scholar
JSTOR b28189796
11 Titles.xlsx

aRel 04
0038-1861
Soundings (New Haven, Conn.)
JSTOR b28190646
26 Titles.xlsx

bRel 03
1549-9804
Journal of the National Association of Biblical Instructors.
JSTOR b25882892
14 Titles.xlsx

bRel 04
0885-2758
Journal of Bible and religion.
JSTOR b23261936
24 Titles.xlsx

bRel 05
0002-7189
Journal of the American Academy of Religion
JSTOR b21673378
52 Titles.xlsx

bRel 05s01 – supplement
0146-9215
Journal of the American Academy of Religion. Supplement.
JSTOR 0
0 Titles.xlsx

bRel 05s02 – supplement
0735-6919
JAAR thematic studies.
JSTOR 0
1 Titles.xlsx: in list for 0002-7189 (see above)

Same list of JSTOR related titles including records found and actions taken.

aRel 01
2163-7652
American college bulletin
JSTOR b28189759
0 Titles.xlsx
11 OCLC (4 lack ISSN)

Print:

1 preferred PCC
0 duplicate
3 lang records (3 lack ISSN).

Film umi:
1 preferred PCC
0 duplicate
0 lang record

Fiche: 0

Elec:
1 preferred PCC
1 duplicate
4 lang records (1 lacks ISSN)

aRel 02
1550-574X
Christian education (Chicago, Ill.)
JSTOR b28190786
7 Titles.xlsx
17 Titles OCLC (8 lack ISSN)

Print:
1 preferred PCC
3 duplicate (3 lack ISSN)
3 lang records (1 lacks ISSN)

Film umi:
1 preferred PCC
1 duplicate (1 lacks ISSN)

Fiche: 0

Elec:
1 preferred PCC
3 duplicate (2 lack ISSN)
4 lang records (1 lacks ISSN)

aRel 03
0361-8234
Christian scholar
JSTOR b28189796
11 Titles.xlsx
18 OCLC (6 lack ISSN)

Print:
1 preferred PCC
2 duplicate (1 lacks ISSN)
2 duplicate aacr (both considered duplicate of later preferred aacr record) (1 lacks ISSN)
2 lang records

Film umi:
1 preferred M Level

2 duplicate (2 lack ISSN)

Fiche: 0

Elec:

1 preferred PCC
2 duplicate
5 lang record (2 lack ISSN)

aRel 04
0038-1861
Soundings (New Haven, Conn.)
JSTOR b28190646
26 Titles.xlsx
32 OCLC (2 lack ISSN)

Print:

1 preferred PCC
1 preferred aacr full level
8 duplicate
1 aacr duplicate
5 lang record

Film umi:

1 preferred full level
1 duplicate (1 lacks ISSN)

Fiche umi:

1 preferred full level

Elec:

1 preferred PCC
1 preferred aacr min level
5 duplicate
6 lang record (1 lacks ISSN)

bRel 03 (parent: aRel 02)
1549-9804
Journal of the National Association of Biblical Instructors.
JSTOR b25882892
14 Titles.xlsx
24 OCLC (9 lack ISSN)

Print:

1 preferred PCC
2 duplicate (1 lacks ISSN)
6 lang records (2 lack ISSN)

Reprint (print):

1 preferred full level (1 lacks ISSN)

Film atla:

1 preferred PCC

Film umi:
1 preferred full level (1 lacks ISSN)
Fiche: 0
Elec:
1 preferred PCC
3 duplicate (1 lacks ISSN)
8 lang record (3 lack ISSN)

bRel 04
0885-2758
Journal of Bible and religion.
JSTOR b23261936
24 Titles.xlsx
38 OCLC records (11 lack ISSN)
Print:
1 preferred PCC
1 preferred full level (reprint)
7 duplicate (4 lack ISSN)
11 lang record (2 lack ISSN)
Film atla:
1 preferred PCC
Film umi:
1 preferred full level (1 lacks ISSN)
1 duplicate
Film unknown:
1 preferred min level of reprint (this was a problem record – probably should have been coded as form “r” maybe a duplicate of preferred reprint record above?)
Fiche aar:
1 preferred full level record
Elec:
1 preferred PCC
3 duplicate (1 lacks ISSN)
9 lang records (3 lack ISSN)

bRel 05
0002-7189
Journal of the American Academy of Religion
JSTOR b21673378
52 Titles.xlsx
78 OCLC records (13 lack ISSN)
Print:
1 preferred PCC
1 preferred aacr PCC
11 duplicates (1 lacks ISSN)

4 aacr duplicates (2 lack ISSN)
18 lang records (1 lacks ISSN)

Film umi:
1 preferred min level
1 preferred full level aacr (1 lacks ISSN)
1 aacr duplicate

Fiche aar:
1 preferred min level (1 lacks ISSN) (added ISSN)
5 duplicates (5 lack ISSN) (including 1 aacr for lack of ISSN & other reasons)

Fiche atlan:
1 preferred full level record (Atlantic Microfilm Corp.)

Fiche nma:
1 preferred min level (National Micrographics Association)
1 duplicate

Fiche sch:
1 preferred full level (Scholars Press)
1 duplicate

Fiche umi:
1 preferred full level
1 duplicate

Fiche unknown:
1 "problem" record: lacks micropublisher: reported to UChicago (1 lacks ISSN):
Reply received: Scholars Press (sch) – not a duplicate
1 lang record

Elec:
1 preferred PCC
1 preferred aacr min level
10 duplicates
1 aacr duplicate (1 lacks ISSN)
12 lang records

bRel 05s01 – supplement
0146-9215
Journal of the American Academy of Religion. Supplement.
JSTOR 0
0 Titles.xlsx
7 OCLC records (2 lack ISSN)
Print:
1 preferred PCC
1 duplicate
2 lang records
Film: 0
Fiche sch:
1 preferred full level

2 problem records (records for single issues. No action taken. (2 lack ISSN)

Elec: 0

bRel 05s02 – supplement
0735-6919
JAAR thematic studies.
JSTOR 0
1 Titles.xlsx
3 OCLC records

Print:

1 preferred PCC
2 lang records

Film: 0
Fiche: 0
Elec: 0

De-Duping status: As of 5-1-2018, no duplicate records merged as yet, based on spot check of three.
De-Duping status: As of 5-29-2018, no duplicate records merged as yet, based on spot check of three.