Notes on Operations

Using Automation and Batch Processing to Remediate Duplicate Series Data in a Shared Bibliographic Catalog

Elaine Dong, Margaret Anne Glerum, and Ethan Fenichel

The application of divergent local practices in a shared bibliographic database can result in unexpected display issues that adversely affect user experience. This is especially problematic when merging databases from multiple institutions accustomed to adopting local practices for their own constituents. The authors describe their experience with the application of automation tools, such as MarcEdit, Excel, and Python, during a large-scale remediation project. They used these tools to analyze, compare, and batch process bibliographic records to remediate obsolete and redundant series data in their shared bibliographic database.

Along with accuracy and comprehensiveness, consistency in cataloging practice improves discovery and identification of resources. Conversely, varying cataloging practice, whether due to local needs or changes to national standards, can result in inconsistent data within a shared bibliographic catalog. The consolidation of bibliographic databases in library consortia may exacerbate these inconsistencies. To maintain metadata quality and update older data to newer standards, catalogers can build on their traditional knowledge and also use data analysis, scripting, and batch manipulation when performing large-scale remediation.

The authors are catalogers at institutions comprising the State University Libraries (SUL) of Florida. As members of the Bibliographic Control and Discovery Subcommittee of the Council of State University Libraries, they formed the Multiple-Series Cleanup Task Force. The Task Force members were chosen due to their complementary skill sets. Two of the members have extensive experience and training in cataloging practice while the third had substantial experience with databases and systems technology before a career in librarianship. One of the members had experience developing Python scripts as a content systems analyst at a financial information provider. Another member has experience with developing XSLT and JavaScript programs. Although these tools were not used for this remediation project, experience with programing language provided a conceptual understanding that assisted with interpreting the Python scripts. All the members had varying experience with data analysis, batch processing, and batch loading as part of their assignments. To aid in these efforts, they independently learned to use MarcEdit through trial and error, webinars, and from peers. Similarly, they also learned how to take advantage of Excel's powerful data analysis tools.

The Task Force was charged with creating a plan to remediate duplicate series data that were causing issues in the catalog's discovery tool. To fulfill its charge, the Task Force identified records in SUL's shared bibliographic database that included obsolete and duplicate series fields that caused display problems.

Elaine Dong (edong@fiu.edu) is the Database & Metadata Management Librarian at Florida International University. **Margaret Anne Glerum** (aglerum@fsu.edu) is the Associate University Librarian and Head of Complex Cataloging at Florida State University Libraries in Tallahassee. **Ethan Fenichel** (fenichele@fau.edu) is the E-Resources Access Management Librarian at Florida Atlantic University.

Manuscript submitted July 11, 2016; returned to authors for revision October 7, 2016; revised manuscript submitted December 5, 2016; accepted for publication March 29, 2017.

The Task Force first analyzed the records using MarcEdit and Excel, and then developed a Python script to compare a subset of the records in the shared bibliographic database of the SULs—known as the Shared Bib—with their corresponding OCLC master records. They ultimately updated the problematic Shared Bib records using a locally developed batch-loading tool. The application of these automation tools saved a significant amount of time rather than manually updating each record. The workflows and processes used for this project serve as an example for how catalogers can approach future remediation projects in an efficient and effective manner.

Literature Review

How best to incorporate quality bibliographic description into a library's catalog has been a topic of discussion in literature for decades.[1] In 2008, *Cataloging and Classification Quarterly* devoted an entire special issue to the topic.[2] High-quality bibliographic description is generally defined as accurate, usable, complete, and consistent.[3] These components are needed for a positive impact on the user experience. Petrucciani writes about the need for consistency and accuracy as prerequisites for establishing trust among the users that the catalog will provide "clear and effective navigation functions among controlled bibliographic entities."[4] Dunsire states, "The efficiency and effectiveness of any information retrieval service requires coherency and consistency in metadata."[5] Harmon acknowledges the direct relationship between the presence of information in the bibliographic record and the library users' retrieval of that record in the discovery interface, and asserts that it is the cataloger's responsibility to support the organization's public service mission in providing access to research materials.[6] Among the key findings of the 2009 OCLC Report, *Online Catalogs: What Users and Librarians Want*, is that "appropriate, accurate and reliable data elements . . . are critical" in retrieving bibliographic descriptions and that "search results must be relevant and the relevance must be obvious."[7] It is that last statement that directly relates to the issues outlined in this paper—multiple series statements and access points are coded to display only in particular local discovery tools, leaving users wondering why the record was retrieved when it does not display the search terms entered.

To maintain the desired quality in their bibliographic database, libraries can outsource their database maintenance, provide it in-house, or use a combination of both. Guajardo and Carlstone describe a Resource Description and Access (RDA) conversion project at the University of Houston Libraries using Marcive, a bibliographic services company, plus in-house staff, to update their catalog records to the new standard.[8] Williams describes an authority remediation project provided by Marcive, followed by subsequent review by the London School of Economics Library staff.[9] Similarly, Finn described an authority control workflow at Virginia Tech that began with an updated authority file provided by Library Technologies Incorporated (LTI), followed by staff using MarcEdit, a free database maintenance program developed by Terry Reese, to edit the authority fields of vendor records before batch loading them.[10] Park and Panchyshyn discuss how they contracted with Backstage Library Works to enrich their MARC records with RDA elements while staff used MarcEdit in-house to create AACR2-RDA hybrid records during Kent State University Libraries' database enrichment project.[11]

Outsourcing database remediation was not an option for the Multiple Series Cleanup Project, so it was performed solely by the Task Force, drawing on earlier projects. Draper and Lederer at Colorado State University Libraries discuss a project using MarcEdit to generate particular field and subfield counts in a set of MARC records in preparation for batch loading. At the University of Minnesota Libraries, Traill and Genereux explained how they transformed Microsoft Office Excel spreadsheets into MARC records using MarcEdit.[12] Sanchez et al. at the Alkek Library at Texas State University-San Marcos described methods using MarcEdit and both Excel and Microsoft Office Word to provide quality control for vendor-supplied records.[13] Myntti and Neatrour demonstrated how MarcEdit and OpenRefine, a free, open-source program, were used to scrub and transform data to update the controlled vocabulary of existing data and to further enrich the metadata with Uniform Resource Identifier (URI) values in preparation for linked data capabilities at the University of Utah.[14]

When there is sufficient in-house expertise, computer programs can be developed for bibliographic database analysis and processing. Myntti and Cothran developed a process to achieve automated authority control for metadata in the University of Utah's digital collection. This process adapted existing services provided by Backstage Library Works to utilize algorithms for reconciling uncontrolled names and subject terms in XML data and replace them with authorized constructions.[15] Frank outlined a method of batch-processing MARC records using MarcEdit and Python, an open source programming language, plus PyMARC, a Python library for parsing MARC record data.[16] To automate the importing of metadata and content during a data migration into the DSpace archive directory format, Walsh at Ohio State University Libraries used Excel, Python, and Perl, another open-source programming language.[17] Mitchell and McCallum explored computational techniques for migrating metadata using OpenRefine and Python.[18] Mitchell later studied data analysis techniques for comparing different library holdings using Python, PyMARC, MySQL, and command line scripts.[19] For the Dewey Decimal Classification

Number "clean-up" at the Library of the Pontifical University Santa Croce, Bargioni et al. shared that seven different Perl programs were developed for queries via the API for their open source ILS, Koha.[20]

SUL Shared Bibliographic Database Overview

SUL members use Ex Libris' Aleph as their integrated library system. In June 2012, the eleven SUL members, in collaboration with the Florida Virtual Campus (FLVC), merged their twenty-three million bibliographic records from separate databases into the Shared Bib of about eleven million records.

SUL members have used OCLC records and vendor records for more than forty years, during which cataloging rules and practices have changed. Part of the need for the Multiple Series Cleanup Project stemmed from the 2008 change when the MARC 440 field (Series Statement/Added Entry-Title) was made obsolete.[21] Another key development was in June 2006, when the Library of Congress (LC) stopped creating authorized series access points (formerly referred to as headings) in conjunction with the transcribed series statement, a practice known as tracing, on its newly created bibliographic records.[22] An untraced series is indicated in the MARC 490 field with a first indicator "0" (490 0_). However, Program for Cooperative Cataloging (PCC) participants and other libraries continued to trace series. In MARC, traced series are encoded as MARC 490 field with a first indicator "1" (490 1_), which indicates that there is a corresponding authorized series access point in a MARC 80X-83X 8xx field.[23]

Before the Shared Bib merge, some SUL members imported different versions of the same OCLC or vendor supplied record, which contained variants in common fields. SUL members also added fields for local data specific to the items at their institution. During the merge, multiple copies of a bibliographic record were combined into one. Due to the difficulty of identifying the particular local data, it was agreed to that all the varying forms of fields would be retained. The subfield $5 was established to label fields with potentially local data. As a result, repeated fields with variations were added to Shared Bib records, including series fields that repeated due to the slight variations of the transcription, incorrect subfield coding, or varying tracing practices. The authors requested a report from FLVC that identified 209,671 records with multiple series MARC fields (440s and 490s).

SUL members share a statewide union discovery layer named Mango, which was developed by FLVC's predecessor the Florida Center for Library Automation (FCLA). Several institutions use a local instance of Mango in addition to the union version for statewide access.[24] To control the display of institution-specific data, FLVC configured Mango to use the SUL members' OCLC MARC Organization Code in MARC subfield $5.

Subsequent to the merge process, the subfield $5 protected fields from being overwritten during the updating of a Shared Bib record with an OCLC master record. Since many fields marked with subfield $5 are not necessarily local data, FLVC later changed the function of subfield $5 to only control display and not to protect the field. The SUL members identified thirty-four fields to protect, irrespective of a subfield $5, since those fields would be likely to contain local data.[25]

Display Issues in the Discovery Layer

The multiple functionalities and extensive use of the subfield $5 resulted in several problems in the Mango discovery layer. The Task Force focused on these issues affecting series data:

1. If a MARC 440 or 490 field includes a subfield $5, that field's series data will display only in the local Mangos corresponding to the MARC organization codes. Figures 1 and 2 show that the University of West Florida (UWF) and the University of North Florida (UNF) Mangos display only the series statements that have MARC 490 fields with the MARC Organization Code for its library, FPeU and FJUNF respectively.
2. If a MARC 440 or 490 field includes a subfield $5, that field will not display in the Union Mango nor in any other local Mango that lacks a corresponding subfield $5 code. Figure 3 shows that the Union Mango does not display any series statements because every MARC 490 has a subfield $5, yet series access points found in the MARC 830 fields do display because they lack subfield $5.
3. Due to the legacy functionality of MARC, Mango treats the MARC 440 as a series access point. If both MARC 440 and 830 are present, both fields display in the local Mango, even if those fields have the same text string. Figure 4 shows that since the 490 fields do not include subfield $5 with the MARC Organization Code for its library (FTS), the USF catalog does not display any series statements. However, since the 440 fields do have subfield $5 FTS, the University of South Florida (USF) catalog displays series access points found in both MARC 440 and 830 fields.

The following screenshots are various displays of the same Shared Bib MARC record containing these series statement and access point fields.

=440 0_ $a Essays in history, economics, & social science, $v 8 $5 FTS
=440 0_ $a Burt Franklin research & source works series, $v 163 $5 FTS
=490 0_ $a Burt Franklin research & source works series #163 $5 FPeU
=490 1_ $a Essays in history, economics, & social science #8. $5 FJUNF $5 FPeU
=830 _0 $a Burt Franklin research & source works series $v no. 163
=830 _0 $a Selected essays in history, economics, & social science, $v 8.

Series note:	Burt Franklin research & source works series #163 Essays in history, economics, & social science #8.
Series:	Burt Franklin research & source works series no. 163 Selected essays in history, economics, & social science, 8.

Figure 1. UWF Mango Catalog

Series note:	Essays in history, economics, & social science #8.
Series:	Burt Franklin research & source works series no. 163 Selected essays in history, economics, & social science, 8.

Figure 2. UNF Catalog

Series:	Burt Franklin research & source works series no. 163 Selected essays in history, economics, & social science, 8.

Figure 3. Union Mango Catalog

Series:	Burt Franklin research & source works series no. 163 Burt Franklin research & source works series, 163 Essays in history, economics, & social science, 8 Selected essays in history, economics, & social science, 8.

Figure 4. USF Mango Catalog

Shared Bibliographic Record Issues

In establishing best practices, SUL members understood that a Shared Bib record should represent a single manifestation, therefore the series statements should not differ among SUL members. However, consortial guidelines allowed for different tracing practices. Please see example 1 below for a case where member institutions chose different tracing practices.

When updating a Shared Bib record, obsolete MARC 440 fields should be replaced with a MARC 490 and its corresponding MARC 830 authorized access point. As discussed, this is a challenge when the MARC 440 fields are indicated as being specific to one of the SUL members (see example 2 below). As discussed in the previous section, the ambiguity around which fields truly are specific to one of the SUL members largely stems from the Shared Bib merge. Example 3 illustrates how this can make cataloging practice more difficult.

Example 1: Multiple MARC 490 fields for different tracing practice on a Shared Bib Record

=001 020001295
=035 __$a(OCoLC)49356140
=440 _0$a**Explorations** in sociology$vv.62$5FTS
=**490 0**_$a**Explorations** in sociology $vv.62$5FBoU$5FU
=**490 1**_$a**Explorations** in sociology ;$v62$5FTaFA$5FMFIU$5FTaSU
=830 _0$a**Explorations** in sociology ;$vv. 62.

The corresponding OCLC record

=001 ocm49356140\
=003 OCoLC
=**490 1**_$a**Explorations** in sociology ;$vv. 62
=**830** _0$a**Explorations** in sociology ;$vv. 62.

Example 2: Obsolete MARC 440 fields on a Shared Bib Record

=001 020000022
=035 __$a(OCoLC)00000069
=**440** _0$aReprints of economic classics **$5FMFIU$5FU**
=**440** _4$aThe Adam Smith library**$5FMFIU**
=490 0_$aReprints of economic classics$5FJUNF$5FTaFA$5FPeU
=490 0_$aThe Adam Smith library $5FJUNF$5FTaFA$5FPeU$5FU

Example 3: Multiple MARC 490 and 830 fields with same tracing practice on a Shared Bib Record

=001 020000093
=035 __$a(OCoLC)00000311
=**490** 1_$aBollingen series, 35:10. The A. W. Mellon lectures in the fine arts**$5FTaSU**
=**490** 1_$aBollingen series, 35. The A. W. Mellon lectures in the fine arts, 10 **$5FSsNC$5FMFIU$5FJUNF$5FPeU$5FBoU$5FTaFA$5FTS$5FU**
=**490** 1_$aBollingen series, 35:10**$5FOFT**
=**490** 1_$aBollingen series, 35. The A. W. Mellon lectures in the fine arts,$v10**$5FFmFGC**
=**830** _0$aBollingen series,$v35.
=**830** _0$aA.W. Mellon lectures in the fine arts ;$v10.

=**830** _0$aA.W. Mellon lectures in the fine arts.
=**830** _4$aThe A. W. Mellon lectures in the fine arts ;$v1961
=**830** _4$aThe A. W. Mellon lectures in the fine arts,$v10

The corresponding OCLC record

=001 ocm00000311\
=003 OCoLC
=**490** 1_$aBollingen series, 35. The A.W. Mellon lectures in the fine arts, 10
=**830** _0$aBollingen series ;$v35.
=**830** _0$aA.W. Mellon lectures in the fine arts ;$v10.

Project Goal

The project's goal was to resolve the issues affecting the display of series data in both the local and the Union Mango while preserving any data specific to each institution. The Task Force devised an automated resolution due to the sheer number of records with problematic attributes. After examining some of these problematic Shared Bib records, the Task Force found that most of the records originated from OCLC. SUL members have relied on OCLC, the largest bibliographic database in the world, as a main source for importing and updating local bibliographic records, even predating the creation of the Shared Bib. Accordingly, the Task Force discovered that most of these problematic local bibliographic records were imported from OCLC a long time ago and have not been updated since.

The example records displayed in the preceding section illustrate problematic Shared Bib records that were no longer compliant with current standards. Most corresponding OCLC master records had since been updated and contained only accurate series pairs. In contrast, the local records contained various forms of series fields that had been contributed over time in each library's individual catalog. These various forms of series fields were then merged into a single Shared Bib record. In addition to correct series data, the OCLC records contained enhancements contributed by OCLC members plus the automatic maintenance performed by OCLC over the years, such as RDA updates and FAST headings. For an example of a full record in Shared Bib compared to its corresponding OCLC record, see appendix A.

The Task Force determined that the best way to update these problematic Shared Bib records would be to overlay them with their latest OCLC master records. This would correct the specific problems with the series data with the added benefits of updating other fields in the local records, including RDA enhancements and additional access points. The Task Force also needed to identify which records were acceptable for overlay and to protect local data. The following section describes the analytical method and the tools used to achieve this goal.

Analysis of Shared Bib and OCLC Records

To identify which Shared Bib records were candidates for overlay, the Task Force performed the following analysis:

1. Shared Bib Records: MARC 035 Field Analysis

The MARC 035 field contains the system control number for the Shared Bib records. The purpose of the MARC 035 field analysis was to identify the locally held records that originated from OCLC and represent the same manifestation compared to those that were provided by other vendors or derived from OCLC records for different manifestations. Examples of the last case were the Shared Bib records in formats different from the corresponding OCLC records.

The authors created a random sample of 1,000 Shared Bib records from the report of 209,671 problematic records.[26] After extracting the MARC records from Shared Bib, the Task Force used MarcEdit to extract just the MARC 035 fields and to copy and paste the results into Excel. The values were sorted and the data were separated into the following four groups:

1. Records with OCLC numbers only (674 records, 67 percent)
2. Records containing more than one MARC 035 field where one of the MARC 035 field values is an OCLC number and another is a vendor identifier (63 records, 6 percent). The majority of these records were identified as vendor records. A separate remediation project is currently underway to address this type of record.
3. ProQuest CIS microfiche records in the Shared Bib with both a MARC 035 field containing an OCLC number and a MARC 035 field containing a proprietary ProQuest number.[27] Some OCLC numbers end with an "x" on the end (36 records, 3 percent).These Shared Bib records are used for microforms and were created by ProQuest from print format OCLC records. These records should not be overlaid by their corresponding OCLC records.
4. Vendor records lacking an OCLC number in MARC 035 fields (285 records, 28 percent). These records could not be updated by the overlaying method since they did not have OCLC records.

After discussion, the Task Force agreed that records in Group 2-4 were not suitable for overlay.

2. Shared Bib Records: Format Analysis

The Task Force developed a Python script to return the necessary MARC data to examine the records in Group 1.[28] In particular, the Task Force focused on the record format as determined by the fixed fields, mainly the MARC 008 field. Using the script, they identified 7,535 records matching Group 1 parameters from 10,000 records that were drawn from the original problem set. The distribution of the formats is shown in table 1.

Table 1 shows that the majority of Group 1 records (89 percent) are print format. There are also small percentages of electronic (5 percent), microform (6 percent), and unknown format (0.3 percent). The Task Force spot checked records for each format and determined that each format needed to be treated differently.

A portion of Shared Bib records coded as microform contained MARC 035 or MARC 019 fields matching OCLC records encoded as print format. In light of this finding, the Task Force added a comparison of the formats of the OCLC records and Shared Bib records as part of the automated analysis. They also determined that records with mismatched formats were not suitable candidates for overlay.

The Task Force determined that records coded as electronic format were not candidates for overlay. The provider-neutral cataloging policy that the PCC implemented in 2009 led to provider-specific records for electronic resources being merged into single provider-neutral records in OCLC.[29] This policy raised concerns about the consistency and comprehensiveness of description in the OCLC master records relative to local records. Before automating the overlay of records for electronic resources, the Task Force wanted to apply additional rigor to the analysis. To complete an iteration of enhancement without resolving this problem, the Task Force simply decided to exclude this category of records.

After reviewing the Shared Bib records with programmatically undetermined formats, the Task Force discovered that they were mostly map or GIS format records. They agreed that these and the Shared Bib records coded as print format were candidates for overlay.

Table 1. Format of Group 1 Records

No. of records with a MARC 035 field beginning with (OCoLC) prefix only	7,535	Percentage
Format: print	6,697	89.0
Format: electronic	391	5.0
Format: microform	422	6.0
Format: unknown	25	0.3

3. OCLC Master Records: MARC 490 and 830 Field Analysis

At this stage of analysis, the Task Force wanted to ensure that any potentially local series data in the Shared Bib would not be lost during the overlay process. To accommodate local practices, they wanted to avoid reversing the tracing of the series in the Shared Bib if the series was not traced on its corresponding OCLC record.

Among the 7,535 OCLC master records corresponding to the Group 1 records that were still candidates for overlay, the Task Force identified eighty-three OCLC bibliographic records (1 percent) that lacked any MARC 490 or 830 fields. Since their Shared Bib records contained MARC 440, 490, or 830 fields, which might be local series, the Task Force agreed that these records were not candidates for overlay. Instead, they created a set of records to be reviewed for authority control by a separate team.

The Task Force also identified 1,222 OCLC bibliographic records (16 percent) that contained MARC 490 0_. They discovered that some of the corresponding series authority records included a MARC 645 subfield $a with a value of "n" (untraced), subfield $5 DLC, and were created before 1989, hence the series were correctly untraced in the OCLC bibliographic records according to LC and PCC standards in Section Z1 of the Descriptive Cataloging Manual.[30] However, some series statements should be changed to traced (MARC 490 1_ and 830 _0 combination) since their series access points were established and should be traced according to their MARC 645 subfield $a with a value of "t" (traced). After discussion, the Task Force decided that these 1,222 records (16 percent) should be parsed for authority review and were not pursued as candidates for overlay.

Unprotected Local Series Data and Access Points in the Shared Bib

Focusing on the preservation of unprotected local series data and access points, the authors collaborated with SUL representatives and colleagues to collect information about data created by each library. This information helped in developing the Python script and determining the best method to identify and protect local data from overlay. The Task Force identified the following three types of local data created by SUL members that was not in the thirty-four protected fields:

1. Access Points for Local Collections

Some SUL members create access points that are only related to materials in their libraries, such as names of specific collections. Since these access points are relevant to a single institution, they do not need to be established

in any authority files. The purpose of these locally created access points is for easy retrieval of associated bibliographic records when these phrases do not appear on the materials. After the Shared Bib merge, SUL members have adopted the use of MARC 79X and 89X fields for these locally created access points.[31] For example, Florida International University (FIU) created and added a MARC 899 field with the phrase "George Wise Collection" to all the bibliographic records for materials donated by George Wise.[32] Before the Shared Bib merge, SUL members used other MARC fields for these locally made access points, including MARC 710, 490, and 830. Below is an example of an access point for a local collection in a MARC 710 field. In Shared Bib, MARC 79X and 89X are among the thirty-four fields protected from OCLC overlay. However, MARC 710, 490, and 830 are not. To preserve data in these fields, the Task Force agreed that the 710 fields should be protected during the overlay process.

=001 020173100
=035 __$a(OCoLC)09370337
=490 0_$aBulletin / Department of Agriculture [new series] ;$vno. 188$5FTaSU
=490 1_$aBulletin / Department of Agriculture, State of Florida ;$v[new ser.], no. 188$5FU$5FTS
=**710 2_$aFloridiana Collection.$5FTS**
=830 _0$aBulletin (Florida. Department of Agriculture) ;$vno.188.
=830 _0$aBulletin (Florida. Department of Agriculture) ;$vnew ser., no. 188.

2. Local Tracing Practices for Series-like Phrases

Some SUL members preferred to trace series-like phrases so that they are indexed and searchable as series and title in Aleph, whereas their SARs in the national authority file instruct catalogers to use the series-like phrases as quoted notes only. For example, authority record number (ARN) 5234175 "Black circle book," a SAR established in the national authority file, instructs catalogers to use the title as a quoted note only. Table 2 below shows the difference between the local and national SAR:

Some SUL members have added the series-like phrases to MARC 490 and 899 fields in Shared Bib bibliographic records as shown in the following example:

=001 032057831
=035 __$a(OCoLC)00289583
=**490** 1_$aA black circle book$5FTaSU
=**899** _0$aBlack circle book$5FTaSU

The local practice of adding the series-like phrase in the indexed MARC 490 and 899 fields will not be found in the corresponding OCLC record. As shown below, the OCLC record uses the series-like phrase as a MARC 500 quoted note only.

Table 2. Series Authority Record for "Black circle book" in Local and National Authority File

SAR in Local Authority Database	SAR in LC/NACO Authority File
=040 \\$aFNP$cFNP	=010 no 00040240
=130 \0$**aBlack circle book**	=040 \\$aNcU$beng$cNcU
=643 \\$aNew York$bGrove Press	=130 \0$aBlack circle book
=644 \\$af$5FJUNF	=643 \\$aNew York$bGrove Press
=645 \\$at$5FJUNF	=667 \\$**aGive phrase as quoted note**.
=646 \\$as$5FJUNF	

=001 289583
=003 OCoLC
=**500** __$a"a black circle book."

To retain data from these local tracing practices, MARC 490, 830 fields in Shared Bib records would need to be compared to the corresponding fields in OCLC master bibliographic records prior to the remediation process to determine if those fields contain local data.

3. Locally Created Series Authority Records

Prior to the SUL members' participation in the Library of Congress (LC)/Name Authority Cooperative Program (NACO), locally created authority records, including those for series headings, existed in each SUL member's local databases. In Florida, the University of Florida (UF), Florida International University (FIU), Florida State University (FSU), and University of North Florida (UNF) libraries are the earliest NACO contributors. NACO participants contribute authority records for names, uniform titles, and series headings to the LC/NACO Name Authority File (NAF). In October 2008, seven libraries, including five university libraries (UF, UNF, FIU, FSU, and USF), two college libraries, and one public library in Florida joined the Florida NACO Funnel. A UF librarian served as the funnel coordinator. This joint endeavor consolidated members' efforts to make a larger contribution to the national authority file and has improved the quality of authority records originating in Florida.[33]

After the Shared Bib merge, all of the locally created authority records were migrated to a combined local authority file in Shared Bib. The Task Force examined a sample of locally created SARs and found that many of them were established in the LC/NACO NAF. The comparison between SARs created locally and those in the national file showed that most of them have the same form of authorized access point (MARC 130 field), while some provided

a different treatment (e.g. Analyzed versus Not analyzed, Traced versus Not traced, Classified as a collection versus Classified separately).[34] These locally created series were added to MARC 440/490/830 fields on Shared Bib records; here is an example:

Locally Created Series in MARC 440 and 490 field on a Shared Bib record

=001 020001980
=**440** _0$aAddison-Wesley series in metallurgy and materials$5FMFIU$5FTS$5FTaSU
=**490** 0_$aAddison-Wesley series in metallurgy and materials$5FJUNF$5FBoU$5FU

The locally created series with unestablished SARs in the national authority file would need to be identified and retained during the overlay process.

Project Workflow and Implementation

Based on the findings from record analysis and information collected about local series practice, the Task Force developed an initial remediation plan. After testing the first 10,000 problem records, analyzing the test results, and adjusting the program logic, the Task Force finalized the workflow (see figure 5). For an account of the project's timeline, please see appendix B.

The Task Force took the following steps to remediate the problematic records:

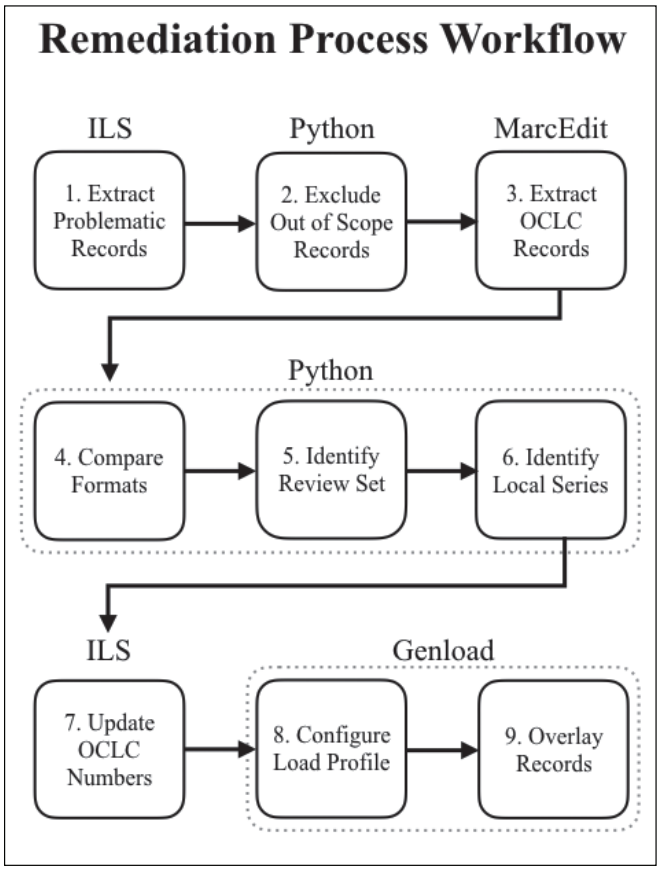

Figure 5. Overall Workflow

Step 1. Use Aleph Services to Extract Problematic Shared Bib Records

The Task Force first extracted the 222,404 problematic MARC records from the Shared Bib in twenty-three batches using a function for record retrieval native to the consortium's cataloging system, Aleph.

Step 2. Use Python Script to Remove Records beyond Scope of Analysis

In the section *Analysis of Shared Bib and OCLC Records*, it is established that when updating Shared Bib records, the Task Force wanted to overlay only non-electronic resource records that could be firmly established as OCLC records. To do this, they collaboratively created a Python script to identify records that originated from OCLC defined by having only "OCoLC" in the MARC 035 prefix. The script also classified the record formats to filter out electronic resource records. After completion of this step, there were 130,692 Shared Bib records remaining.

Step 3. Use MarcEdit to Extract OCLC Master Records

Using the MarcEdit Z39.50 Client's batch processing function, the Task Force retrieved the corresponding OCLC master records.

Step 4. Use Python Script to Compare Formats of Shared Bib and OCLC Records

Following the download, the Task Force developed a second Python script to compare the format of the Shared Bib records with their corresponding OCLC master records (see appendix C). Record pairs with mismatched formats were identified and excluded.

Step 5. Use Python Script to Identify OCLC Records for Authority Review

The Task Force used the same script for format comparisons to build the Authority Review Set. This set would contain the Shared Bib records whose corresponding OCLC records that either lacked a MARC 490 field or contained a MARC

490 0_. These records would not be considered for overlay but instead were referred to a separate team to analyze for compliance with consortial authority policies. After this process was applied to all twenty-three batches, the Task Force identified 25,951 OCLC records and corresponding Shared Bib records as the Authority Review Set.

Step 6. Use Python Script to Identify Local Series

After excluding records with mismatched formats and records from the Authority Review Set, the script performed a text string comparison between the series data on the Shared Bib records and those on the corresponding OCLC master records (see appendix D). The goal of the comparison was to identify local series on the Shared Bib records and to flag them for the "Do Not Overlay" set. The records with matching series data were placed in the "Suggest Overlay" set.

To eliminate non-critical mismatches between text strings, the Task Force added additional rules to normalize the data prior to the comparison process. The Task Force chose to remove "his," "her," or "him" from the beginning of the series text string because the words were used inconsistently, especially in Shared Bib records. The differences were not critical enough to classify as a mismatch. The Task Force chose to remove numbers from subfields $a and $p since the series numbering had been incorrectly entered in these subfields. Diacritics were normalized so that differences in character encodings did not result in a mismatch.[35] All of the text normalization rules applied in the script are listed below.

- Strip out the following data in subfield $a and subfield $p for MARC 440, 490, and 830 fields before comparison:
  - Initial articles in English, French, and Spanish: the, a, an, el, los, la, las, un, unos, una, unas, le, la, l', les, un, une, des
  - His, Her, Him
  - Punctuation marks including ‘ ’ “ ” ... ! : ; , . [ ] < > ( ) { } - | / \
  - Numbers
  - Volume and number abbreviations ("NO" "V" "VOL")
- Additional text manipulation
  - Convert all text to uppercase
  - Normalize text encoding of diacritical marks to use UTF-8

Comparison Logic

After the script normalized the series data in the form of text strings, it performed a series of comparisons. The order of the comparisons was significant and in each comparison, either the text strings were considered as equal or the Shared Bib record would not be considered a candidate for overlay and was flagged for the "Do Not Overlay" set. In each comparison, only the subfields $a and $p were used from the MARC fields 440, 490 and 830.

First, the script compared all of the MARC 440 fields from a Shared Bib record with the MARC 490 and 830 fields of its corresponding OCLC master record. If the script determined that the series data did not match, the Shared Bib record was placed in the "Do Not Overlay" set. If the Shared Bib MARC 440 matched the OCLC master record's series data, the script proceeded to the next step.

Second, the script compared all of the MARC 490 fields from the Shared Bib record with its corresponding OCLC master record's series data. If the script determined the series data did not match, the Shared Bib record was placed in the "Do Not Overlay" set. If the Shared Bib MARC 490 matched the OCLC master record's series data, the script proceeded to the next step.

In the third and final comparison, the script compared all of the MARC 830 fields from the Shared Bib record with its corresponding OCLC master record's MARC 830 data. If the script determined the series data did not match, the Shared Bib record was placed in the "Do Not Overlay" set. If the Shared Bib MARC 830 matched the OCLC master record series data, the script added the Shared Bib record to the "Suggest Overlay" set. It would then repeat the comparisons for the next Shared Bib record in the batch. In total, by using the script, the Task Force placed 53,802 records in the "Suggest Overlay" set. For diagrams of steps 5 and 6, see figure 6.

Step 7. Use Aleph Services to Update OCLC Numbers of Shared Bib Records

While performing the comparisons in step 3, the script identified 243 cases where the Shared Bib record's OCLC number in the MARC 035 field did not match any OCLC Master record due to a merge of OCLC records. To accurately update the Shared Bib records, the Task Force first updated the MARC 035 field value to match the current OCLC number. The Task Force completed this using an automated service native to Aleph. If the current OCLC record was also in the Shared Bib, the Task Force deleted the duplicate record.

Step 8. Use GenLoad Profile to Protect Local Fields from Overlay

GenLoad is a record loading utility created by FLVC for SUL members to load MARC data into the Shared Bib.[36] GenLoad performs each load based on the profile configuration.

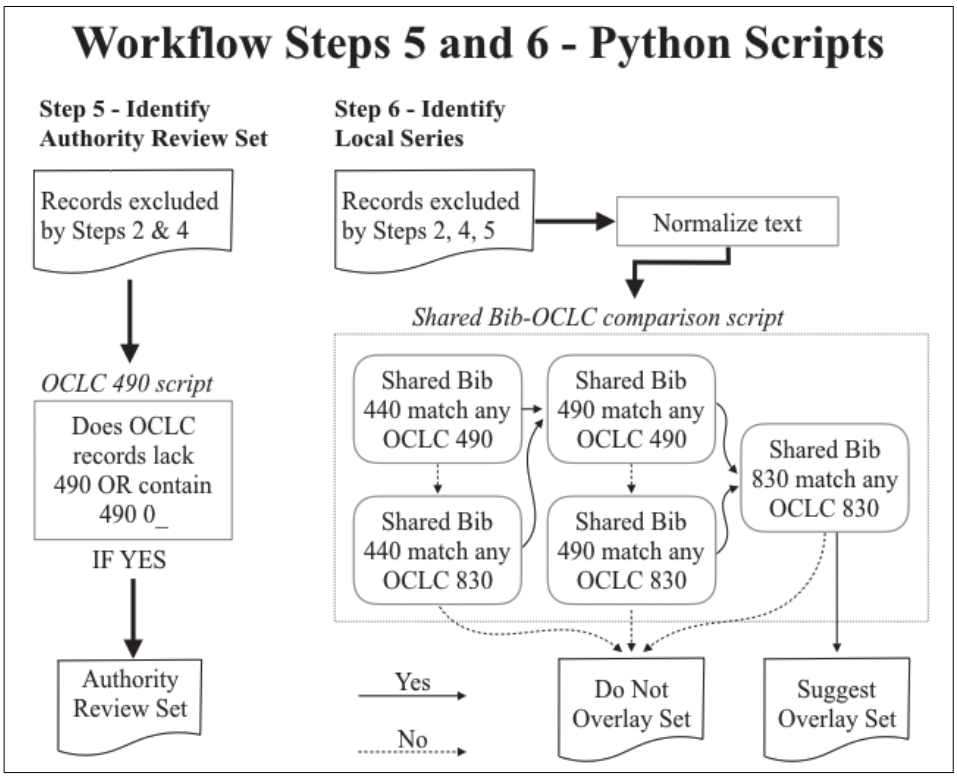

Figure 6. Expanded Workflow, Steps 5–6

The profile controls which data to insert and which data to protect or replace. The Task Force created a custom configuration to protect local data during the overlay process.

The Task Force added all of the fields to be protected to the GenLoad Profile. These include the thirty-four fields established by FLVC. Among these thirty-four fields, there are two non-standard MARC fields. The first is LKR, which is used to link bibliographic records in the Shared Bib. The second is TKR, which is a pre-merge holdover used to create an indexed string on the bibliographic records. Beyond the thirty-four fields, the Task Force included another non-standard MARC field to indicate record status, abbreviated as "STA." For example, "STA $aProvisional" on a Shared Bib record indicates it's a provisional record. The Task Force also included MARC 520, 599, and 710 fields, because they are likely to contain local data.

The profile protects local data in the following fields:

1. Established MARC fields to protect: LKR, TKR, 351, 500, 501, 506, 520, 533, 540, 541, 542, 545, 561, 562, 563, 583, 584, 590, 690, 691, 699, 790, 791, 797, 845, 896, 897, 898, 899, 909, 951, 970, 655 _7 with the following subfield $2: rbprov, rbbin, rbgenr, rbpap, rbpri, rbpub, rbtyp,
2. Added MARC fields: STA, 520, 599, 710, 655 _7 $2 local

Step 9. Used GenLoad to Batch Overlay Shared Bib Records with OCLC Records

Following a review period in which other SUL members provided feedback, the Task Force proceeded to the final step. They downloaded the OCLC master records that corresponded to the Shared Bib records in the "Suggest Overlay" set using MarcEdit. Following initial testing, the Task Force used GenLoad to overlay 51,818 Shared Bib records within two weeks.

Results

The Task Force's analysis and the resulting procedure that they developed culminated in the identification of 53,802 records as candidates for overlay, including approximately 2,000 duplicates from the originally identified 222,404 records with multiple series issues. Following the Task Force's work, a total of 51,818 Shared Bib records were overlaid. See appendix E for an example of a Shared Bib record before and after the overlay process.

This duplicate series data remediation project has made significant improvements in the quality of the Shared Bib database. The updates to series fields improved presentation, retrieval, and access for users of the consortial discovery systems. The project has also impacted the Shared Bib environment for internal maintenance. Concurrently, SUL members were preparing to merge their database with the State College libraries, as part of a migration to a new Next-Generation Integrated Library System. The improvements have reduced the overall amount of work required to complete that migration.

Future Projects Possibilities

The steps taken to remediate series data in our shared bibliographic database utilizing OCLC master records demonstrates a process that is repeatable and expandable. In our project, the use of PyMarc allowed us to create a customized process for analyzing and manipulating a large amount of MARC data. The writing of scripts by members of a cataloging team opens the possibilities for new procedures. Cataloging units could replicate the process to remediate

MARC fields containing local data by distinguishing locally created data in local records from non-local data in OCLC master records. As part of the remediation, the units could move local data to appropriate locally defined fields, such as MARC 89x fields.

A future project expanding on this work would allow bibliographic records to receive automated quality checks. Scripts could identify problems in local records, errors in OCLC records prior to loading into a local database, or perform comparisons between local and OCLC records. Resolutions for identified problems would follow either through further scripts or human intervention. By building the scripting into workflows such as database maintenance, copy cataloging, and batch loading, bibliographic records are reviewed automatically for predictable problems.

In the future, a shift to a linked data bibliographic environment will reduce the need for this process. The procedure relies on the model of a bibliographic record in the database as a document. Shifting bibliographic description to records as data graphs, or serialized data, will remove the need to analyze the full bibliographic document since data will be updated at a more granular level.[37] This question remains to be addressed as the structures and models of linked library data are developed.[38] The expansion of the scripting abilities in cataloging units is likely to be an essential component in the transition to the new models and workflows.

Conclusion

When large-scale changes to library bibliographic data are required, cataloging departments may lack the resources to suspend other projects and will spend hours manually updating records. By exploiting new technologies and skills, they can quickly adapt their data to the latest systems, cataloging standards, and changes in practice. The ability to utilize automated tools to analyze and batch process data is now an essential skill for librarians responsible for bibliographic data.

SUL faced large-scale changes that began with a system migration and were exacerbated by revisions to the practices of recording series data. When it became apparent that the existing practices were adversely affecting users, the Task Force identified how to bulk update series data. By using generally available tools—Microsoft Excel, MarcEdit, Python, and a locally developed data loader, GenLoad—the Task Force eased the analysis and largely automated the record update process. Three tech savvy catalogers completed this work without the involvement of formal software developers or systems experts. The Task Force made significant updates to the Shared Bib environment for all SUL members, with minimal help. In doing so, they demonstrated the value of leveraging automation in consortial collaboration.

While updating the Shared Bib records with OCLC master records, the Task Force made improvements beyond the series data that were initially the target for enhancement. In many cases, bibliographic records in the Shared Bib had not been updated in a long time. The latest versions of the OCLC master records contained improved description that would not have been captured through normal workflow processes. For example, the updated OCLC records contained the results of OCLC automated enhancements and authority control such as RDA updates and FAST subject headings.

The Task Force's analysis helped highlight the benefits of establishing best practices between SUL members. Accordingly, the Task Force made recommendations for SUL members on how to transcribe series in general and to add local series. One issue that the Task Force encountered was different tracing practices among individual libraries in a shared database. The Shared Bib Guidelines that all SUL members follow state that individual libraries may apply varying practices for analysis, tracing, and classification practice found in the LC Authority File. The Task Force recommended that the best practice is to use the OCLC master record's treatment of the series fields rather than alter the Shared Bib record. If the OCLC bibliographic or authority record needs to be revised to the current standard, that should also be done. If the library initiating the change is not authorized to edit the OCLC record, they can contact an SUL member who is authorized to do so.

Another issue that the Task Force observed is that individual libraries have used different fields for local series before Shared Bib. After the records have been merged into a single database, it takes a significant effort to identify and protect local data from being overlaid and causes serious challenges for data remediation. The authors feel that in a shared database, it is better to put local series and other local data into actual locally defined fields such as the MARC 590, 69X (local subject access fields), 79X (local added entry fields), 89X (local series added entries), and 9XX (local data elements) and minimize the use of other fields for local data. If a future library system allows it, it would be ideal to record local data in a separate section (e.g., holding records), not in bibliographic records, which would make the management and maintenance of the shared database much easier and efficient.

References and Notes

1. Janet Swan Hill, "Is It Worth It? Management Issues Related to Database Quality," *Cataloging & Classification Quarterly* 46, no. 1 (2008): 5–26, https://doi.org/10.1080/01639370802182885; Barbara Schultz-Jones et al., "Historical and Current Implications of Cataloguing Quality for Next-Generation Catalogues," *Library Trends* 61, no. 1 (2012): 49–82.

2. "Special Issue: Bibliographic Database Quality," *Cataloging & Classification Quarterly* 46, no. 1 (2008).
3. Hill, "Is It Worth It?"; Peter S. Graham, "Quality in Cataloging: Making Distinctions," *Journal of Academic Librarianship* 16, no. 4 (1990): 213–18; Heather Moulaison Sandy and Felicity Dykas, "High-Quality Metadata and Repository Staffing: Perceptions of United States-Based OpenDOAR Participants," *Cataloging & Classification Quarterly* 54, no. 2 (2016): 101–16, https://doi.org/10.1080/01639374.201.1116480; Philip Hider and Kah-Ching Tan, "Constructing Record Quality Measures Based on Catalog Use," *Cataloging & Classification Quarterly* 46, no. 4 (2008): 338–61.
4. Alberto Petrucciani, "Quality of Library Catalogs and Value of (Good) Catalogs," *Cataloging & Classification Quarterly* 53, no. 3–4 (2015): 303–13.
5. Gordon Dunsire, "Collecting Metadata from Institutional Repositories," *OCLC Systems & Services* 24, no. 1 (2008): 51–58.
6. Joseph C. Harmon, "The Death of Quality Cataloging: Does It Make a Difference for Library Users?," *Journal of Academic Librarianship* 22, no. 4 (1996): 306–7.
7. Karen S. Calhoun et al., "Online Catalogs: What Users and Librarians Want: An OCLC Report" (Dublin: OCLC, 2009), accessed December 4, 2016, http://www.oclc.org/content/dam/oclc/reports/onlinecatalogs/fullreport.pdf.
8. "Marcive," accessed December 4, 2016, http://home.marcive.com; Richard Guajardo and Jamie Carlstone, "Converting Your E-Resource Records to RDA," *Serials Librarian* 68, no. 1–4 (2015): 197–204.
9. Helen K. R. Williams, "Cleaning up the Catalogue," *Library & Information Update* (2010): 46–48.
10. Library Technologies, Inc. "We are the Authority Control Specialists," accessed December 4, 2016, https://www.authoritycontrol.com; Mary Finn, "Batch-Load Authority Control Cleanup Using MarcEdit and LTI," *Technical Services Quarterly* 26, no. 1 (2009): 44–50; "About MarcEdit," accessed December 4, 2016, http://marcedit.reeset.net/about-marcedit.
11. Amey L. Park and Roman S. Panchyshyn, "The Path to an RDA Hybridized Catalog: Lessons from the Kent State University Libraries' RDA Enrichment Project," *Cataloging & Classification Quarterly* 54, no. 1 (2016): 39–59, http://www.tandfonline.com/doi/abs/10.1080/01639374.2015.1105897.
12. Daniel Draper and Naomi Lederer, "Analysis of Readex's Serial Set MARC Records: Improving the Data for the Library Catalog," *Government Information Quarterly* 30, no. 1 (2013): 87–98, https://doi.org/10.1016/j.giq.2012.06.010; Stacie A. Traill and Cecilia Genereux, "Strategies for Catalog Management of Electronic Monographs in Series," *Serials Librarian* 65, no. 2 (2013): 167–80.
13. Elaine Sanchez et al., "Cleanup of NetLibrary Cataloging Records: A Methodical Front-End Process," *Technical Services Quarterly* 23, no. 4 (2006): 51–71, https://doi.org/10.1300/J124v23n04_04.
14. Jeremy Myntti and Anna Neatrour, "Use Existing Data First: Reconcile Metadata before Creating New Controlled Vocabularies," *Journal of Library Metadata* 15, no. 3–4 (2015): 191–207, https://doi.org/10.1080/19386389.2015.1099989; "OpenRefine," accessed December 4, 2016, http://openrefine.org.
15. Jeremy Myntti and Nate Cothran, "Authority Control in a Digital Repository: Preparing for Linked Data," *Journal of Library Metadata* 13, no. 2-3 (2013): 95–113; "Backstage Library Works," accessed December 4, 2016, http://www.bslw.com/about/.
16. Heidi Frank, "Augmenting the Cataloger's Bag of Tricks: Using MarcEdit, Python, and PyMARC for Batch-Processing MARC Records Generated From the Archivists' Toolkit," *Code4lib Journal* 20 (2013), accessed December 4, 2016, http://journal.code4lib.org/articles/8336; "About Python™: Python.org," accessed December 4, 2016, http://www.python.org/about/; "PyMARC," accessed December 4, 2016, http://pymarc.sourceforge.net.
17. Maureen P. Walsh, "Batch Loading Collections into DSpace: Using Perl Scripts for Automation and Quality Control," *Information Technology & Libraries* 29, no. 3 (2010): 117–27; "About Perl: www.perl.org," accessed December 4, 2016, http://www.perl.org/about.html; "DSpace: DSpace is a turnkey institutional repository application," accessed December 4, 2016, http://dspace.org.
18. Erik Mitchell and Carolyn McCallum, "Old Data, New Scheme: An Exploration of Metadata Migration using Expert-Guided Computational Techniques," *Proceedings of the American Society for Information Science and Technology* 49, no. 1 (2012) 1–10, https://doi.org/10.1002/meet.14504901091.
19. Erik T. Mitchell, "Reconciling Holdings Across Multiple Libraries: A Study in Data Analysis Techniques," *Technical Services Quarterly* 33, no. 2 (2016): 154-169, https://doi.org/10.1080/07317131.2016.1135000.
20. Stefano Bargioni et al., "Obtaining the Dewey Decimal Classification Number from Other Databases: a Catalog Clean-up Project," *Italian Journal of Library & Information Science* 4, no. 2 (2013): 176, https://doi.org/10.4403/jlis.it-8766.
21. "MARC Proposal No. 2008-07," accessed December 4, 2016, http://www.loc.gov/marc/marbi/2008/2008-07.html.
22. "Series at the Library of Congress: June 1, 2006" (Washington, DC: Library of Congress, 2006), accessed December 4, 2016, http://www.loc.gov/catdir/cpso/series.html.
23. "MARC 21 Bibliographic: 80X-83X - Series Added Entry Fields" (Washington, DC: Library of Congress, 2008), accessed December 4, 2016, http://www.loc.gov/marc/bibliographic/bd80x83x.html.
24. "Florida Academic Library Services Cooperative (FALSC): Discovery Tools" (Tallahassee, Florida: FALSC, 2015), accessed December 4, 2016, https://libraries.flvc.org/discovery-tools.

25. Bibliographic Control and Discovery Subcommittee of the Council of State University Libraries, "Shared Bib Guidelines. Appendix IIIA: Fields to Protect on Overlay from OCLC Gateway Import," (Gainesville, FL), accessed December 4, 2016, https://sharedbib.pubwiki.fcla.edu/wiki/index.php/Shared_Bib_Guidelines_Online#APPENDIX_IIIA:_Fields_to_Protect_on_Overlay_from_OCLC_Gateway_Import.
26. A VBA macro was created using MOD(ROW(),209)=1 to select every 209th row.
27. "Legislative Publications: CIS Congressional Bills, Resolutions & Laws on Microfiche (1933–2008)" (Ann Arbor, Michigan: ProQuest, 2014), accessed December 4, 2016, http://cisupa.proquest.com/ws_display.asp?filter=cis_leaf&item_id={1D481C6F-CA7A-4929-B4B0-BC90D20FAC71}.
28. Ethan Fenichel, "FLVC_490_Duplicates," GitHub, accessed December 4, 2016, https://goo.gl/ttjBtp.
29. "Program for Cooperative Cataloging (PCC) Provider-Neutral E-Resource MARC Record Guidelines," accessed December 4, 2016, http://www.loc.gov/aba/pcc/scs/documents/PCC-PN-guidelines.html.
30. "Descriptive Cataloging Manual Section Z1 and LC Guidelines Supplement to MARC 21 Format for Authority Data" (Washington, DC: Library of Congress, 2016), accessed December 4, 2016, http://www.loc.gov/catdir/cpso/z1andlcguidelines.html.
31. Bibliographic Control and Discovery Subcommittee of the Council of State University Libraries, "Shared Bib Guidelines. Section 3.4.8: Local Series" (Gainesville, FL), accessed December 4, 2016, https://sharedbib.pubwiki.fcla.edu/wiki/index.php/Shared_Bib_Guidelines_Online#3.4.8_Local_Series.
32. "899 Local Series Added Entry-Uniform Title" (Dublin, Ohio: OCLC, 2016), accessed December 4, 2016, https://www.oclc.org/bibformats/en/8xx/899.html.
33. Priscilla William, "TSPC Authorities Subcommittee Report on the Florida NACO Funnel" (Gainesville, FL, 2010), accessed December 4, 2016, http://csul.net/sites/csul.fcla.edu/uploads/authorities-NACOrpt-11-03-10.pdf.
34. "MARC 21 Format for Authority Data—64X Series Treatment General Information" (Washington, D.C.: Library of Congress, 2008), accessed December 4, 2016, www.loc.gov/marc/authority/ad64x.html.
35. Geoffrey Spear, "More Unicode Issues - Diacritics This Time" [Online forum comment]. Aug.10, 2015, Message posted to https://groups.google.com/forum/#!msg/pymarc/w9iy9dTb5xQ/RcbEg4VaHQAJ.
36. "GenLoad" (Gainesville, FL: Florida Virtual Campus, 2012), accessed December 4, 2016, https://support.flvc.org/knowledge-base/kbdw/KBA-01484-R4V7.
37. "RDF 1.1 Concepts and Abstract Syntax," accessed December 4, 2016, http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.
38. "BIBFRAME Training at the Library of Congress" (Washington, DC: Library of Congress, 2016), accessed December 4, 2016, http://www.loc.gov/catworkshop/bibframe/.

Appendix A. A Full Record in Shared Bib and its Corresponding OCLC Record

Shared Bib Record:

=LDR 04517cam a22008534a 4500
=001 020001295
=005 20111213154942.0
=008 020313s2002\\\\enka\\\\\b\\\\001\0\eng\\
=010 \\$a2002024878
=015 \\$aGBA2-54901
=019 \\$a50433750$a51052019$a51681752
=020 \\$a0333984994 (alk. paper)
=020 \\$a0333984994
=035 \\$a(OCoLC)49356140
=040 \\$aDLC$beng$cDLC$dUKM$dTJC$dMUQ$dNLGGC$dBAKER$dBTCTA
$dYDXCP$dOCLCG$dIG#$dKAAUA$dGEBAY$dOCLCQ$dFUG
=650 \7$aTi=042 \\$apcc
=050 00$aHM656$b.S63 2002
=082 00$a304.2/3$221
=084 \\$a71.02$2bcl
=245 00$aSocial conceptions of time :$bstructure and process in work and everyday life /$cedited by Graham Crow and Sue Heath.
=260 \\$aHoundmills, Basingstoke, Hampshire ;$aNew York :$bPalgrave MacMillan,$c2002.

=300 \\$axvii, 266 p. :$bill. ;$c23 cm.
=440 \0$aExplorations in sociology$vv.62$5FTS
=490 0\$aExplorations in sociology$vv.62$5FBoU$5FU
=490 1\$aExplorations in sociology ;$v62$5FTaFA$5FMFIU$5FTaSU
=504 \\$aIncludes bibliographical references (p. 247-263) and index.
=650 \0$aTime$xPsychological aspects.
=650 \0$aTime$xSocial aspects.
=650 \6$aTemps$xAspect psychologique.
=650 \6$aTemps$xAspect social.
=650 07$aZeit.$2swd
=650 07$aAlltag.$2swd
=650 07$aZeitwahrnehmung.$2swd
=650 07$aAufsatzsammlung.$2swd
=650 17$aSociologische aspecten.$2gtt
=650 17$aPsychologische aspecten.$2gtt
=650 17$aTijd.$2gtt
=700 1\$aHeath, Sue.
=700 1\$aCrow, Graham.
=700 1\$aHeath, Sue,$d1964-
=830 \0$aExplorations in sociology ;$vv. 62.

Its Corresponding OCLC Record:

=LDR 03318cam a22007574a 4500
=001 ocm49356140\
=003 OCoLC
=005 20150914140806.0
=008 020313s2002\\\\enka\\\\\b\\\\001\0\eng\\
=010 \\$a 2002024878
=040 \\$aDLC$beng$cDLC$dUKM$dTJC$dMUQ$dNLGGC$dBAKER$dBTCTA
$dYDXCP$dOCLCG$dIG#$dKAAUA$dGEBAY$dOCLCQ$dOCLCF$dOCLCO$dOCLCQ
=015 \\$aGBA254901$2bnb
=019 \\$a50433750$a51052019$a51681752
=020 \\$a0333984994$q(alk. paper)
=020 \\$a9780333984994$q(alk. paper)
=035 \\$a(OCoLC)49356140$z(OCoLC)50433750$z(OCoLC)51052019$z(OCoLC)51681752
=042 \\$apcc
=050 00$aHM656$b.S63 2002
=082 00$a304.2/3$221
=084 \\$a71.02$2bcl
=245 00$aSocial conceptions of time :$bstructure and process in work and everyday life /$cedited by Graham Crow and Sue Heath.
=260 \\$aHoundmills, Basingstoke, Hampshire ;$aNew York :$bPalgrave MacMillan,$c2002.
=300 \\$axvii, 266 pages :$billustrations ;$c23 cm.
=336 \\$atext$btxt$2rdacontent
=337 \\$aunmediatedbn2rdamedia
=338 \\$avolumebnc2rdacarrier
=490 1\$aExplorations in sociology ;$vv. 62
=504 \\$aIncludes bibliographical references (pages 247-263) and index.
=650 \0$aTime$xSocial aspects.
=650 \0$aTime$xPsychological aspects.
=650 \6$aTemps$xAspect social.
=650 \6$aTemps$xAspect psychologique.

```
=650  \7$aTime$xPsychological aspects.$2fast$0(OCoLC)fst01151056
=650  \7$aTime$xSocial aspects.$2fast$0(OCoLC)fst01151066
=650  17$aTijd.$2gtt
=650  17$aPsychologische aspecten.$2gtt
=650  17$aSociologische aspecten.$2gtt
=650  07$aZeit.$2swd
=650  07$aAlltag.$2swd
=650  07$aZeitwahrnehmung.$2swd
=650  07$aAufsatzsammlung.$2swd
=700  1\$aCrow, Graham.
=700  1\$aHeath, Sue,$d1964-
=830  \0$aExplorations in sociology ;$vv. 62.
```

Appendix B. Project Implementation Timeline

Jan. 2015	A report of 209,671 Shared Bib records with multiple series (MARC 440/490/830) fields was generated by FLVC
Mid-April	Multiple-Series Cleanup Task Force was formed to analyze potential solutions for the issues resulting from the multiple series in these records
May-Aug.	Task Force analyzed sample records and began fact-finding
June	Task Force developed strategy: use Python program to flag records that contain local data, and use GenLoad to batch overlay records with obsolete and duplicate series using their corresponding OCLC master records
June-Aug.	Task Force developed, tested and finalized the Python scripts
Last week of Aug.	Task Force configured and tested GenLoad profile for loading OCLC master records. Following the successful test loads, FLVC approved the GenLoad profile.
Sep. 3	Task Force requested and received an updated report from FLVC that included 222,404 Shared Bib records with multiple series.
Sep.	Task Force executed the Python script against the new report resulting in the identification of the following: • 53,802 records in the Suggest Overlay Set • 106 duplicate records from Suggest Overlay Set which were sent for deduplication • 243 Shared Bib records whose OCLC number needed to be updated due to merge of OCLC master records
Oct.	Task Force presented the project at the Council of State University Libraries (CSUL) Cataloging, Authorities and Metadata Committee (CAM) and the FLVC Members Council on Library Services Technical Services Standing Committee (TSSC) meeting. This also began the review period where the Task Force solicited feedback prior to any additional updates.
First two weeks of Nov.	Task Force batch loaded the OCLC master records from Suggest Overlay Set to update problematic records in Shared Bib

Appendix C. Python Script for Format Determination

```
# called from main script, to get format information
# Return Formats
# lFormat = returnFormat(lDict[key])
# lDict[key] is the dictionary of fields for a given MARC record
# mFormat = returnFormat(mDict[aDict[key]])

def returnFormat(dict):
    # extract the code from the 008 23 values
    formatCode = 'None'
    for tag in dict['fields']:
        for k in tag:
            if k == '008':
```

```
            formatCode = tag[k][23:24]
    format = ''
    if formatCode in ['s', 'o', 'q']:
        format = 'electronic'
    elif formatCode in [' ', 'r', 'd']:
        format = 'print'
    elif formatCode in ['a', 'b', 'c']:
        format = 'microform'
    else:
        format = 'unknown'

    return format
```

Appendix D. Python Script for Identification of Local Series Values

```
# part of the main script - calls to the function that does the comparisons
# l440 is the cleaned series strings from the SharedBib MARC 440
# l490 is the cleaned series strings from the SharedBib MARC 490
# l830 is the cleaned series strings from the SharedBib MARC 830
# m490 is the cleaned series strings from the OCLC MARC 490
# m830 is the cleaned series strings from the OCLC MARC 830

# placeholder list, wasteList allows the procedure to send a value to the function as a placeholder
wasteList = ['-1']

#Compare Local440
compResult = betterComparison(l440, m490, m830, wasteList, wasteList)
# this part follows each comparision (is excluded from below cases)
if len(compResult) > 0:
        sendForLocalCheckResults = [lSysNumber, oclcNumberL, '440', local440]
        writeLocalCheckResults(sendForLocalCheckResults, lSysNumber)
        logString = logString+'\n\tComparison Strings Not Found (440):'+'\n\t\t'+compResultString
        writeBibsForOverlay(lSysNumber, oclcNumberL, '0')
        logResult(str(keyCounter), logString)
        keyCounter += 1
        continue

#Compare Local490
compResult = betterComparison(l490, m490, m830, wasteList, wasteList)

#Compare Local830
compResult = betterComparison(l830, m830, wasteList, wasteList, wasteList)

# the betterComparison function called that actually does the comparisions.
def betterComparison(lista, listb, listc, listd, liste):
    unfoundSeriesStringList = []
    badEndingValues = ['V']
    beginningWords = ['he ', 'her', 'his', 'she']

    listaa = []
    listbb = []
    listcc = []
```

```
    listdd = []
    listee = []

    for a in lista:
        if len(a) > 0 and a.upper()[-1] in badEndingValues:
            a = a[0:len(a)-1].strip()
        listaa.append(a.upper())
    for a in listb:
        if len(a) > 0 and a.upper()[-1] in badEndingValues:
            a = a[0:len(a)-1].strip()
        listbb.append(a.upper())
    for a in listc:
        if len(a) > 0 and a.upper()[-1] in badEndingValues:
            a = a[0:len(a)-1].strip()
        listcc.append(a.upper())
    for a in listd:
        if len(a) > 0 and a.upper()[-1] in badEndingValues:
            a = a[0:len(a)-1].strip()
        listdd.append(a.upper())
    for a in liste:
        if len(a) > 0 and a.upper()[-1] in badEndingValues:
            a = a[0:len(a)-1].strip()
        listee.append(a.upper())

    for series in listaa:
        if series in listbb:
            continue
        if series in listcc:
            continue
        if series in listdd:
            continue
        if series in listee:
            continue

        unfoundSeriesStringList.append(series)

    return unfoundSeriesStringList
```

Appendix E. Example of a Shared Bib Record before and after the Overlay Process

Before

=001 020014504
=035 __$a(OCoLC)00001935
=040 __$aDLC$cDLC$dm.c.$dFNP
=050 0_$aBL1405$b.D4
=050 0_$aBQ1138$b.D4
=090 __$aBL1405$b.D4
=092 __$a294.3$bD286b
=100 1_$aDe Bary, William Theodore,$d1919-$ecomp.
=245 14$aThe Buddhist tradition in India, China & Japan.$cEdited by Wm. Theodore De Bary. With the collaboration of Yoshito Hakeda and Philip Yampolsky and with contributions by A. L. Basham, Leon Hurvitz, and Ryusaku Tsunoda.

=260 __$aNew York,$bModern Library$c[1969]
=300 __$axxii, 417 p.$c20 cm.
=**440** _0$aReadings in Oriental thought**$5FTS$5FTaSU**
=**440** _4$aThe Modern library of the world's best books$v<205>**$5FTS**
=**490** **0**_$aThe Modern library of the world's best books [205]**$5FTaSU$5FPeU$5FU$5FMFIU**
=**490** **0**_$aReadings in Oriental thought**$5FPeU$5FJUNF$5FBoU$5FU$5FMFIU**
=**490** **0**_$aThe Modern library of the world's best books**$5FBoU**
=**490** **1**_$aThe Modern library of the world's best books$v[205]**$5FJUNF**
=504 __$aBibliography: p. [399]-401. Bibliographical footnotes.
=650 _0$aBuddhism$xCollections.
=650 _0$aBuddhism$xSacred books.
=**830** _0$aModern library of the world's best books;$v[205]
=**830** _0$aReadings in Oriental thought.
=**899** _0$aWedig collection.$5FMFIU
=951 __102aFAU01:000255331;$5FBoU
=951 __104aFIU01:001349334;$5FMFIU
=951 __105aFSU01:000547336;$5FTaSU
=951 __109aNFU01:000231416;$5FJUNF
=951 __110aSFU01:000000770;$5FTS
=951 __108aUFU01:000812032;$5FU
=951 __111aWFU01:000222854;$5FPeU

After

As a result of the overlay process, MARC 440, 490, 830, and other fields were updated to reflect the OCLC master record. As an added benefit, the Shared Bib record received the more complete data in the master record including the MARC 33x fields, FAST headings, extra subject access points, MARC 505, and 710 fields were added. Local fields on the Shared Bib record, including MARC 899 and 951 fields, were protected.

=001 020014504
=019 __$a1261666$a462181729$a911553216
=035 __$a(OCoLC)00001935
=040 __$aDLC$beng$cDLC$dOCL$dBTCTA$dITC$dCBC$dHIL$dDEBBG$dOCLCF $dP4I$dOCLCQ$dOCLCO$dTWS$dTAMSA
=050 00$aBQ1138$b.D4
=050 14$aBL1405$b.D4
=100 1_$aDe Bary, William Theodore,$d1919-$ecompiler.
=245 14$aThe Buddhist tradition in India, China & Japan.$cEdited by Wm. Theodore De Bary. With the collaboration of Yoshito Hakeda and Philip Yampolsky and with contributions by A. L. Basham, Leon Hurvitz, and Ryusaku Tsunoda.
=260 __$aNew York,$bModern Library$c[1969]
=300 __$axxii, 417 pages$c20 cm.
=336 __$atext$btxt$2rdacontent
=337 __$aunmediated$bn$2rdamedia
=338 __$avolume$bnc$2rdacarrier
=**490** **1**_$aReadings in Oriental thought
=**490** **1**_$aThe Modern library of the world's best books [205]
=504 __$aIncludes bibliographical references (pages 399-401. Bibliographical footnotes).
=505 0_$aEarly Buddhism -- The life of Buddha as a way of salvation -- "The greater vehicle" of Mahayana Buddhism -- Tantricism and the decline of Buddhism in India -- The coming of Buddhism to China -- The schools of Chinese Buddhism -- The introduction of Buddhism to Japan -- Saicho and the lotus teaching -- Kukai and esoteric Buddhism -- Amida and the pure land -- Nichiren's faith in the lotus -- Zen.
=650 _0$aBuddhism$vSacred books.
=650 _7$aBuddhism.$2fast$0(OCoLC)fst00840028

=650 07$aBuddhismus.$2swd
=651 _7$aChina.$2swd
=651 _7$aIndien.$2swd
=651 _7$aJapan.$2swd
=650 07$aBuddhismus.$0(DE-588)4008690-2$2gnd
=651 _7$aChina.$0(DE-588)4009937-4$2gnd
=651 _7$aIndien.$0(DE-588)4026722-2$2gnd
=651 _7$aJapan.$0(DE-588)4028495-5$2gnd
=655 _7$aCollections.$2fast$0(OCoLC)fst01424032
=710 2_$aRogers D. Spotswood Collection.$5TxSaTAM
=776 08$iOnline version:$aDe Bary, William Theodore, 1919-$tBuddhist tradition in India, China & Japan.$dNew York, Modern Library [1969]$w(OCoLC)610373932
=**830** _0$aModern library of the world›s best books ;$v205.
=**830** _0$aReadings in Oriental thought.
=**899** _0$aWedig collection.$5FMFIU
=951 __102aFAU01:000255331;$5FBoU
=951 __104aFIU01:001349334;$5FMFIU
=951 __105aFSU01:000547336;$5FTaSU
=951 __109aNFU01:000231416;$5FJUNF
=951 __110aSFU01:000000770;$5FTS
=951 __108aUFU01:000812032;$5FU
=951 __111aWFU01:000222854;$5FPeU