2020

Methods for in-sourcing authority control with MarcEdit, SQL, and regular expressions

Mike Monaco, *The University of Akron*

Methods for in-sourcing authority control with MarcEdit, SQL, and regular expressions

Mike Monaco

Coordinator, Cataloging Services

The University of Akron, Akron, Ohio, USA

https://orcid.org/0000-0001-7244-5154

The University of Akron

302 Buchtel Common

Akron, Ohio 44325-1712

Office: 330-972-2446

mmonaco@uakron.edu

***This is an Accepted Manuscript of an article published by Taylor & Francis in the Journal of Library Metada on December 20, 2019, available online: http://www.tandfonline.com/*10.1080/19386389.2019.1703497**

Methods for in-sourcing authority control with MarcEdit, SQL, and regular expressions

ABSTRACT

This is a report on a developing method to automate authority control in-house (that is, without an outside vendor), especially for batch-loaded bibliographic records for electronic resources. A SQL query of the Innovative Sierra database retrieves entries from the "Headings used for the first time" report. These entries are then processed with some regular expression substitutions to create a list of terms suitable for batch searching in the OCLC Connexion client. Two approaches to this problem are described in detail and results compared. A similar method for using the "Unauthorized headings" report from the SirsiDynix Symphony ILS is also described.

Keywords: Authority control, automation, Regular expressions, SQL, batch processes, workflows, Sierra ILS

Shorter title: Methods for in-sourcing authority control

Background

Like many, perhaps most, university libraries in the United States, the majority of The University of Akron University Libraries' collection budget has shifted from print/tangible to electronic resources (e-resources), while cataloging staff time has been reduced through attrition and the assignment of additional non-cataloging responsibilities. Shifting from print resources where manual authority work for individually vetted bibliographic (bib) records is possible to batches of e-resources bib records loaded *en masse* without the same individual vetting has meant changing the approach to authority work for most of the new titles added to the collection. Because outsourced authority vendors are not an option, authority control remains in-house, but by working with the systems unit, the cataloging unit is developing a method to automate much of the authority workload associated with loading large batches of bib records.

Literature review

Authority control is a fundamental and perennial challenge in librarianship, and a great deal has been written about authority control, automation, and challenges posed by vendor-supplied bib records. This review focuses on recent work on (1) how the quality of vendor-supplied batches of bib records has been assessed, (2) attempts to control the quality of such records and the headings in them, and (3) efforts to automate authority work.

The importance of quality metadata for access and discoverability, and the centrality of authority control to quality control in a bibliographic database, is assumed rather than argued for in the present paper. Snow (2017) provides an in-depth review of the literature on the importance of metadata quality.

No discussion of authority control can ignore the use of vendors to "outsource" the drudgery of managing headings and loading authority records (ARs). Park (1992) presents a relatively early look at moving from manual (in-house) authority files to automated authority files maintained by vendors. Tsui & Hinders (1999) look at outsourcing authority work with vendors, which at the time was the only real alternative to finding ARs manually for each access point. Their cost-benefit analysis compared OCLC charges and credits and the associated staff time to the cost of a vendor contract and associated staff time. Aschmann (2002) gives a detailed plan for outsourcing that includes creating an RFP, working with a vendor, and forming an in-house authority control team. Ten years after Park's hopeful outline, Aschmann found that outsourcing did not necessarily save staff time.

Jackson (2003) discusses some of the advantages and perils of automated authority control. While the ability to automatically update headings in the catalog is a great benefit, the downsides include the limits of computer matching (which may produce amusing errors) and the need to manually identify authorities to add to the catalog. Jackson concludes that vendor-supplied authority control carries the same benefits and perils, and hopes that future integrated library systems (ILS) will provide better options for in-house authority control. Velluci (2004) surveys the major vendors of authority control and explains the services they offer. It is an interesting snapshot of the state of authority control in 2004. One of the leading vendors mentioned is no longer in business, and the author notes resistance, on the part of vendors, to the idea of an international authority file. It is an accurate and detailed overview of the services available, excepting only services that have developed since 2004, such as updating bib records to use RDA forms of headings and adding URIs to MARC fields. Zhu & Von Sheggern (2005) outline various quality control services that can be provided by vendors and try to set realistic

expectations for librarians about these services with examples of what can and cannot be automated. This article serves as an excellent overview of how automated authority control works, explaining normalization, matching, and which elements of authority and bib records are typically utilized. The authors also provide a checklist of common options and questions that libraries should ask of vendors. Williams (2010) gives a case study of database cleanup with Marcive, noting some of the limits of automation identified by Zhu & Von Sheggern (2005) and some local issues such as the strain loading batches of bib records put on their ILS, and how they dealt with those challenges.

Some recent work has taken up the idea of managing large scale quality control and authority work projects internally. Kreyche, Lisius, & Park (2010) describe a process at Kent State University Libraries for updating name ARs with death dates added since the NACO policy change which made adding death dates more routine. While it is an important example of a large-scale in-house project, the vendor Backstage Library Works eventually began offering the same service. Ziso, LeVan, & Morgan (2010) describe a method that, rather than using ARs within the database, queries OCLC's WorldCat Identities file to direct users to authorized access points and related works for searches. It is certainly an outside-the-box approach, but most library catalogs still rely on internal authority files, and the method does not help update headings in bib records, so bib record quality would need to be addressed by other workflows. Mak (2013) gives a detailed look at a process at Michigan State University Libraries to cope with the mass re-issue of name authority records (NARs) by the Library of Congress, when many NARs were revised to meet RDA standards. Mak describes a process where ARs in the local catalog are exported, converted to a format allowing the extraction of control numbers for batch searching, and then comparing the retrieved ARs to the exported ARs to identify and select

updated records to load. An AutoIT script automates most of this this process and even updates bib records within the ILS. Cook (2014) provides a roundup of useful tools for manipulating metadata, including programs, development environments, and programming languages that can be used to manipulate MARC records. Some of these tools were utilized in the present project.

Carrasco, Serrano, & Castillo-Buergo (2016) describe a tool for matching headings in the context of a large database with possible duplicates. Their work is notable for relying on bib records to disambiguate names. This was accomplished by analyzing time periods and dates in the bib records associated with names rather than using ARs. Dong, Glerum, & Fenichel (2017) describes a process for resolving a problem that was more or less unique to their shared database: duplicate series data. This is useful to other libraries because the authors detail their planning process and practical lessons learned. The article also includes a good literature review on database quality and describes some approaches and projects in large-scale authority control undertaken elsewhere. Wolf (2019) describes processes that use existing lists of changed or updated ARs (the Library of Congress "Weekly Lists" and OCLC's "Closed Dates in Authority Records") to extract record identifier numbers, and then queries these numbers in the local Sierra ILS (hereafter, Sierra) to determine which ARs need to be re-loaded into Sierra. Wolf's process involves using regular expressions to extract the relevant data, JSON queries of the Sierra database, and batch searches in the OCLC Connexion Client (Connexion), making her work somewhat similar to the present project. Indeed it is a complimentary project; where the present project begins with internal notifications (headings reports) Wolf's begins with external notifications (the aforementioned Library of Congress and OCLC lists).

A natural catalyst for searching for large scale authority control solutions has been the increasing practice of batch loading bib records, especially records for intangible e-resources.

The batch loading of e-resource bib records creates several challenges for authority control, as vendor-supplied bib records may be of varying quality and are loaded in volumes that make evaluating the records individually impractical.

Sanchez, Fatout, Howser, & Vance (2006) is one of the first publications to address the use of "non-traditional, non-ILS supplied editing utilities to correct MARC records prior to loading" (p. 54). Their paper describes their use of MarcEdit, Word, and Excel to correct errors in bib records provided by NetLibrary. These corrections were carried out in batches, but authority work was carried out manually by catalogers. While manual authority work on small batches of e-resource bib records was feasible in 2006, the growth of e-resources in library collections and the reduction of library staff renders such an approach less practical today.

Heinrich (2008) details quality enhancements made to electronic book (e-book) bib records both pre- and post-load. The pre-load work included vetting different collections and requesting customizations based on local practices. The post-load work included deduplication of titles, transferring local information to the batch-loaded records, and establishing overlay protection for the local data fields. However no attempt was made at authority control for electronic serials (e-serials) because the "records are unstable" (p. 15) -- that is, e-serial bib records are frequently updated and redistributed, making local changes to records less permanent than they are for e-books. Moreover, the batch-loaded bib records were also excluded from the headings reports out of concern that the large number of records would "overwhelm the capacity of the headings reports (p. 15). Finn (2009) describes pre-load authority work on batches of bib records at Virginia Tech. Their procedures are a mix of outsourced and internal work. First, Library Technologies, Inc. (LTI) edits the batch files to correct certain common errors and creates a report on "unlinked" (that is, uncontrolled) headings. Then library staff use MarcEdit to make

changes to the access points in the bib records based on LTI reports. LTI also supplies ARs for the library to load. Global updates and headings reports in the ILS are used after loading the bib records to cover additional corrections. Martin & Mundle (2010) offer a typology of authority problems (broadly: access issues, load issues, and record quality issues), explain their procedures for dealing with them, and emphasize the usefulness of talking to vendors as a tactic for maintaining quality control. They focus in particular on Springer e-books, a collection that also proved vexing to other consortia. Wu & Mitchell (2010) describe some of the e-book record quality issues the University of Huston Libraries has found and how they are addressed with MarcEdit batch processes, as well as the difficulties posed by changing cataloging standards, particularly the preference for provider-neutral records.

Panchyshyn (2013) introduces a procedure for quality control via a checklist. Authority control is managed at Kent State by isolating batches of e-resource bib records from other records (p. 27-28). Like Heinrich (2008), Panchyshyn warns that the costs associated with of authority control may make it inadvisable for certain kinds of resources -- in this case, e-resources that will not be held "in perpetuity" (p. 34). Beisler & Kurt (2012) describe a task force used to deal with issues with e-resources and batch loading workflows, developing a form similar to Panchyshyn's checklist for managing workflow. They mention very little on quality control and automated authority processing however. David & Thomas (2015) look at the quality of bib records for e-resources. They note that the quality of bib records is especially important for e-resources because these resources cannot be found on the shelf, and user browsing and selection takes place in the catalog, based mainly on the bibliographic metadata displayed there (p. 802). They focus on the types of errors that occur in access points and the time and cost of correcting them. Their study of user searches confirmed that title, author, and subject fields are

the most important access points, both because they are most frequently chosen for single-field searches and because their analysis of keyword searches found that title, author, and subject terms were the three most commonly entered kinds of search terms. Of course all three of these access points are controlled by ARs, further highlighting the importance of authority work for access.

Flynn & Kilkenny (2017) describe dealing with the problem of e-resource bib record quality at the consortia level. Their paper describes the evolving policies and procedures that were put in place to improve record quality in OhioLINK. These are focused on changes to bib records -- some manual and some automated. They also include a helpful review of the literature on vendor record quality and discuss how they worked with various vendors to improve record quality at the source. Van Kleeck, Nakano, Langford, Shelton, Lundgren, & O'Dell (2017) examine record sources, again highlighting the importance of record quality for e-resources. They conclude that OCLC bib records distributed via WorldShare Collection Manager (WCM) are generally equal to or superior to the vendor-supplied records from other sources that they examined. The record sources are identified in this study (as opposed to David & Thomas (2015) and Flynn & Kilkenny (2017) who anonymize the vendors and publishers), making this an especially helpful article for librarians developing their own workflows. Their emphasis on record quality (especially authorized access points) underscores the importance of authority control for e-resources which are primarily accessed through OPACs or discovery layers dependent on these access points. Thompson & Traill (2017) describe a method to check record quality with Python scripts that evaluate quality using a rubric that gives credit for the presence of authorized access points, call numbers, and descriptive fields which affect discovery such as summaries and contents notes. The records' scores according to the rubric are used to separate records that can

be batch loaded from those that will need human intervention to assure completeness and correctness. This project has had the added value of helping compare the relative quality of different sources of bib records, and confirming Van Kleeck et al.'s observation that WCM provides better records than most vendors.

Tingle & Teeter (2018) describe an effort to make e-resources visible in a fairly literal manner. Proxies for titles and topics were placed on the shelves among print resources. The project highlights how significant an issue the discoverability of e-resources remains, but does not particularly address record quality within the catalog.

Automation of authority work at The University of Akron

Even libraries with authority control vendors often find it impractical or cost-ineffective to have authority outsourcing for e-resources bib records. As discussed in the literature review, e-resources pose a particularly vexing problem because the records are often of low quality, because the records are not expected to remain for long or will be updated with new records at regular intervals, and/or because the sheer number of incoming records can be daunting. At The University of Akron (UA), a large public research university which does not use an authority control vendor, a process was developed to leverage some free software, simple database queries, and existing capabilities already present in the ILS and bibliographic utilities to create a process that improves and controls the access points in bib records with minimal staff effort, and retrieve supporting AR in batches. The process is an example of successful collaboration between librarians with different areas of functional expertise and at different institutions, and we hope our initial successes will inspire other librarians to push themselves to develop skills beyond those traditionally employed within their units.

In developing this method to download batches of ARs to support (and update) headings in incoming bib records, the goal is to automate authority control in-house, especially for batch-loaded bib records for e-resources.

Before loading, batches of bib records for e-resources have their access points for names and topics compared to the Library of Congress' Linked Data Service (LDS) via the MarcEdit report "Validate Headings." This report changes headings in bib records that match variant access points for authorities in the LDS to the authorized forms. The bib records are then loaded into Sierra. This triggers headings reports in Sierra. The "Headings used for the first time" report lists entries for headings that are new to the catalog and therefore do not match ARs in the catalog. This report can be queried with SQL to retrieve text strings to search against the authority file in OCLC via a batch process in Connexion and download matching ARs in batches.

An earlier version of the process will also be described, which involved using a text editor to sort access points by type and then run a series of find/replace operations using regular expressions (regexes, singular: regex) to normalize the access points for batch searching. Some pointers for applying the method in the SirsiDynix Symphony ILS follow. Symphony has a different approach to headings reports than Sierra, but Symphony's reports can still yield usable textual search strings if the report output is processed with a series of regexes similar to those used in the Sierra methods.

Statistics collected to track the success rate of the headings validation tool in MarcEdit and the batch searching of ARs based on the SQL queries are provided.

The conclusion assesses the cost in staff time versus benefit in improved access, and discusses the lessons learned by the authors in this collaboration, as well as suggesting possible refinements and improvements of the process and areas for further exploration.

Pre-load authority work with MarcEdit

At UA, several procedures are followed to improve the quality of bib records before loading them into Sierra. There are two categories of procedures: collection-specific tasks and heading validation. Unlike Virginia Tech's procedures as reported in Finn (2009), this work is carried out entirely in-house.

Most e-resource bib record collections have specific sets of edits that are always applied either before loading (in MarcEdit's MarcEditor program) or during the load (with specialized load tables for the collections). These edits may be local customizations (collocation fields to identify the collection, local call numbers and location codes, etc.), or for a few collections they may be more extensive, such as adding form/genre headings to streaming video collections. For a few collections, recurrent errors that have not been adequately addressed by the record suppliers have their own set of tasks in MarcEditor. The most extreme case is a streaming video collection that has recurring errors in access points, such as incorrect forms of names, qualifiers incorrectly added to corporate body headings, and problems with subdivisions coding (missing or improperly coded delimiters and subfield codes). In many cases these edits are made in MarcEdit because the applicable ARs do not have matching variant access points that would enable the ILS to automatically "flip" the access points in the bib records, and because Sierra reports but does not automatically flip variant forms of headings when a bib record is loaded (rather, automated processing is triggered when the ARs are loaded). These pre-load edits make improvements to record quality, but the most dramatic and efficient processing is the second category, utilizing MarcEdit's "Validate Headings" report.

Heading validation in MarcEdit compares access points in bib records to the authorities in the Library of Congress' LDS. Variant headings for names are flipped to the authorized form if there is an exact match. The validate headings report is a routine part of the workflow at UA for many batches of records. Because the validation report provides a statistical log of the changed headings, the statistics of each set processed are compiled to determine the relative quality of records from different publishers and whether the time required to run the report is justifiable. It was determined that there was little benefit from running the report on the brief bib records supplied for the discovery layer, but on the other hand some collections benefited significantly, especially those that had been harvested at some point from the Library of Congress or OCLC and which therefore had older forms of headings. Two years of data collection (March 2017-March 2019) has demonstrated that of a total of 1,230,195 bib records loaded in batches, 32,249 access points were changed from a variant form to the authorized form. Because the Validate Headings report notes headings in 1xx, 6xx, and 7xx fields separately, it was possible to separately track name access points and topical access points. This was helpful as some sets, such as streaming video, tended to have far more name access points than would be typical of e-books or serials. The results for sets of bib records from different vendors were compared to the results for bib records from OCLC (via WorldShare Collection Manager), with the assumption (supported by Flynn & Kilkenny (2017)) that OCLC records were a reasonable benchmark for acceptable record quality. Using this benchmark, UA only continued using the Validate Headings report for sets that had a rate of correction higher than that of the OCLC record sets. Some selected collections are summarized on table 1, Summary of MarcEdit Validate Headings on selected record sets; the OCLC row is bolded for emphasis as it served as the benchmark for deciding whether the time invested in the report was worthwhile.

[place table 1 here]

Pre-load authority work is particularly beneficial to the workflow because Sierra can identify headings that have not been used before in the catalog ("Headings used for the first time") which perforce do not have corresponding ARs in the catalog: an AR's 1xx field would constitute a previous use of the heading. But because many of the headings in the bib records have been validated or changed to match existing ARs, it is more likely that ARs corresponding to the headings can be retrieved. The reported "new" headings are supplied in a report within Sierra, which made the next steps -- automated retrieval of ARs in-house -- possible. Two versions of the method are detailed below, because the two slightly different approaches have different strengths and weaknesses.

Post-load authority work with Headings reports

The remainder of this paper will describe the development and implementation of a method to accomplish authority control by loading ARs matching headings that have been flagged as "new" to the catalog by headings reports in the ILS.

For clarity, the two different versions of the method are referred to as "Alpha" and "Beta." A third process for another ILS is dubbed "Gamma." The method has three components, which will be referred to as a query, processing, and batch searching. The query retrieves data; the processing prepares the data to be batch searched. Batch searching uses the batch processing module in Connexion to retrieve ARs. The query and processing vary in each version of the method, and it is hoped that the discussion of how they developed and how the different methods compare in terms of efficiency and success will be helpful to others adapting the method to their own libraries.

The Alpha method: background and query

The initial project began with the somewhat obvious thought that it would be nice to be able to gather the headings in the "Headings used for the first time" report in Sierra and batch search them in Connexion. See figure 1, Sample "Headings used for the first time" report entry. A cataloger will recognize the MARC field listed as "Field" in the report. Corresponding MARC fields also appear in ARs. The challenge would be collecting the MARC data in a form that could entered into Connexion searches.

[place figure 1 here]

A colleague in Systems (Susan DiRenzo Ashby, Coordinator, Systems, The University of Akron) identified the location of the report's components in Sierra's database, and another colleague (Michael Dowdell, Systems Administrator, The University of Akron) devised a simple SQL query to collect the MARC fields with the triggering headings. pgAdmin is used to run these queries and place the results in a comma-separated values (.csv) file. pgAdmin is a user interface for accessing databases, executing SQL queries, and managing the results. The .csv file, once the contents are processed (normalized to remove MARC and Sierra codes and tags and potential stop words, operators, or commands), can in turn can be entered into Connexion's batch searching tool to retrieve matching ARs. These ARs ultimately are loaded in support of the bib records. Over time, through trial and error and with help from Craig Boman (Discovery Systems Librarian, Miami University) the query was refined.

The Alpha method queries the Sierra database for the terms listed under "Field:" in the report. The SQL query was:

```
SELECT field
FROM sierra_view.catmaint
```

```
WHERE condition_code_num=1
ORDER by field
;
```

The SQL query is asking for a particular column of data (field), in a particular table (sierra_view.catmaint), where another column in the table (condition_code_num) has a particular value (1). This has exactly the desired effect: the query returns the data labeled “Field:” from all entries in the “Headings used for the first time” report. In the case of the entry depicted in figure 1, the data is:

a1001 |aWolfram, Adolph,|earranger of music,|einstrumentalist

Thus, all of the MARC coding (tags, indicators, subfield delimiters) and also the Sierra field group tag (here, the initial “a”) are returned by the query. This data would interfere with a batch search in Connexion, since the Connexion search is querying the WorldCat authority file, which contains only authorized access points and variants. Additional data such as the relator terms in the example (“|earranger of music,|einstrumentalist”) also interfere with searching. This problem is addressed later in the “processing” component of this procedure.

The field group tag is useful as it distinguishes name headings (tagged “a” for “author” or “b” for “other author”) from subject headings (tagged “d”). This is important because the Sierra requires separately loaded ARs for names when they are used as name access points (Sierra tag “a” or “b” and MARC tags 1xx or 7xx) or as subjects (“d” and 6xx). The Connexion batch searching tool on the other hand requires separate searches for topical headings and name headings. Fortunately it is possible to search the index of Library of Congress (LC) names, which includes personal names, corporate bodies, conferences, and uniform titles (including name/title headings). The possible combinations of tags and headings are laid out in table 2,

Headings types in Sierra and WorldCat. The shaded area highlights situations where name headings are used as subject access points.

[place table 2 here]

This is why the SQL script includes the command to ORDER the output by "field". ORDER sorts the data alphanumerically. The fields starting with "a" or "b" will be separated from the "d"s. Furthermore those starting with "d600" through "d630" would all be grouped together, regardless of the order they appeared in the headings report. Sorting the full fields, with the initial Sierra and MARC tags, effectively groups these different uses of headings. That is, the sorted list is ordered into three groups: name headings used as name access points (or "names-as-names"), name headings used as subject access points (or "names-as-subjects," shown shaded in table 2), and subject headings. (A few other field group tags may also appear in the report, depending on the local settings used, but these too would be gathered by tags.) The three types of headings were then manually "cut and pasted" in a text processing application (in this case, Editpad) into three distinct files to be searched and loaded with slightly different criteria: the names-as-names which are searched as LC names and loaded as names authorities, the names-as-subjects which are searched as LC names and loaded as subject headings, and the subject headings which are searched as LC Subject Headings (LCSH) and loaded as subject headings.

The Alpha method: processing

The Connexion batch searching tool can import text lists of terms to search. The problem though remained: how to search just the data in MARC subfields that would be useable in these searches. Returning to the example in Figure 1, the goal is to search just the words "Wolfram" and "Adolph" and not the words "a1001" "|aWolfram," "Adolph,|earranger" "of", and

“music,|einstrumentalist”, which is how Connexion would parse the field as retrieved. The solution arrived at for the Alpha was to use a series of regexes to find and delete the extraneous data, which is mostly readily identified by MARC codes, and also to strip out punctuation and common stop words and operators that would confound the searches. The stop words and operators can appear both in subject and name -- especially name/title or uniform title, headings. Consider headings such as: “Same-sex divorce,” “Actors with disabilities,” “Cyrus, the Great, King of Persia, -530 B.C. or 529 B.C.,” and “Gone with the wind **(**Motion picture **:** 1939**)**”. The underlined words are interpreted by Connexion as potential operators or stop words, and the punctuation is interpreted as syntax for commands, any of which can interfere with keyword search. The stop words slow down the batch process as they are not indexed and waste effort. The operators and command syntax can cause errors that stop the affected searches. Occasionally, some name elements are identical to WorldCat index labels and will not be readily searched as keywords, because the batch process interprets them as commands lacking proper punctuation. For example, the family name “Su” will be interpreted as the label “su” (for the LCSH index of WorldCat’s authorities) and regarded as an error as it is missing the “:” or “=” which would tell Connexion whether it is a keyword or browse search of that index. There is little to be done in such cases, as removing these name elements is unlikely to create a search with just one match. However the stop words and operators can generally be removed with no loss of precision.

A somewhat complicated series of “find/replace” operations using regexes were therefore performed in the separated text files of names and subjects. The complete list of expressions used follow:

1. (.*\|a)

2. (\|db\. ca\. |\|db\. |\|d\. ca\.|\|dd\. |\|dca\. |-ca\. |\|dfl\. ca\. |\|dfl\.)
3. (\|e.*|\|4.*|\|0.*|\|j.*)
4. (\|.)
5. (“|;|:|\(|\)|\?| and | or |&c\.|&| in | an |,| the | for | on | so | with | to | by |”|’| be | that |\.{3}| near | same)

The first expression simply selects everything up to, and including, “|a” which is how Sierra represents “subfield a” in the MARC field. So, for the example from figure 1, this selects “b7001 |a”. This selection is replaced with nothing; that is, it is simply deleted. The other expressions are all replaced with a blank space, so that the remaining terms do not run together. This is important because the Sierra database does not store spaces that appear before or after subfield delimiters in the MARC record.

The second expression selects commonly occurring AACR2 abbreviations that occur in names with uncertain or incomplete dates. These abbreviations are generally selected in the context of a name heading’s subfield d (hence the “\|d” preceding some tokens); other likely contexts are signified in the expression such as “b. ca.,” “-ca.” and so on. These are likely to occur in older record sets which some vendors distribute. They may also exist in older bib records in the catalog and appear in the report because of some other edit that was made to the record. The example in figure 1 does not have any such abbreviations however.

The third expression selects relationship terms and identifiers, again including the subfield delimiters themselves. In figure 1, “|earranger of music,|einstrumentalist” would be selected. The fourth expression selects any remaining subfield delimiters and codes, such as subfield q (marking a fuller form of name). The last expression selects a variety of punctuation marks and common stop-words and operators.

Running these find and replace substitutions is not especially time-consuming, but they must be run in order and require some attention to detail. Figure 2 shows two screenshots of some actual Sierra fields output by the Alpha query. On the left is the raw output, on the right is the same screen after processing.

[place figure 2 here]

Batch searching

At this point the data, saved as a plain text file, can be imported into Connexion for batch searching. Names (whether used as names or subjects in Sierra) should be searched with the default index "LC Names" (nw:) selected; subjects with "LCSH" (su:). The batch was run with a limit of one hit per match. This limitation to a single hit avoids situations where human intervention might be required to decide between two or more similar headings that are partial matches to the entry. Such partial matches might be name/titles that matched just the name, modified name headings that matched an entry with no modifier, and so on. As an example, consider the name "Colombo, Maria" from figure 2. A search of the name authority file for "nw:colombo maria" yields twelve hits for names containing those words, but none exactly match the entry. On closer inspection none can be identified with the Maria Colombo in UA's catalog anyway, but even if one of the multiple hits were a match, there would be no way to automate selection of the correct heading. Moreover, there is a limit to the number of records that can be stored in a Connexion save file (10,000) and including more than one match would potentially fill the save file before all the terms in the batch are searched.

This was the procedure was carried out at UA for four months in 2017, with queries made about once a week. The reports had a mean average of 4412 entries, mostly due to bibliographic

batch loads and a simultaneous project of re-loading certain e-resource collection bib records. About 52% of the entries were names-as-names, 6% names-as-subjects, and 42% subjects. The greatest success by far was had searching the names-as-names. 59% of the name-as-name entries returned a unique AR, while just 23% of names-as subjects and 5% of subjects did the same. The lower success rates for name and topical subject headings can be partly explained by the fact subdivisions were always included in the authority searches, but ARs established with main headings plus subdivisions are relatively rare. Because the local installation of Sierra was not a version that could ignore subdivisions when creating the headings report, only subdivided ARs would match subdivided headings. So, it made sense to try to find ARs that also have the subdivisions.

In August of 2017 the project was put on hold as upgrades to Sierra were planned, and by good fortune another librarian at a conference (Craig Boman, Miami University) suggested a tweak that could eliminate (1) the need to separate the data retrieved in the query and (2) most of the processing.

The Beta method: query

Mr. Boman suggested altering the query use to **SELECT index_entry** rather than **SELECT field**.[1] The "index_entry" is the data labeled "Indexed as Author" (or "Indexed as Subject", etc.) in the heading report. In figure 1. this is simply "adolph wolfram." These index entries are ready to batch search, for the most part. Because the UA implementation of Sierra does not index the title part of name/title headings in the author index, there is less need to remove stop words and operators from the names. Punctuation is not present in the index entries either. But of course there remains the issue of separating names-as-names, names-as-subjects, and subjects. This was

[1] C. Boman (personal communication, May 14, 2018)

accomplished with another tweak. A condition was added to the query, based on the prefixes in the **field**. Instead of running one query and then separating and normalizing the output with the Alpha processing, the separation could be accomplished by running three distinct queries. The resulting data needs less processing, because the MARC coding and punctuation are already absent.

Names-as-names were selected with the following query that exploits regexes in the search. The use of a regex is indicated by the tilde (~) and the expression enclosed in single quotes.

```
SELECT index_entry
FROM sierra_view.catmaint
WHERE condition_code_num=1 and field ~'^a|^b'
;
```

The "WHERE" conditional now focuses on fields that begin with an "a" or "b" -- that is, on fields with the index group tag for "name" (a) or "other name" (b). As mentioned above, the "index_entry' will not contain subfield t, so articles and other stop words and operators are less common. Even so, conference names, place names, and uniform titles may occur in these as "names" or "other names" so there may still be some terms that will confound Connexion batch searches. For example, the abbreviation for Oregon ("Or.") will appear in the index_entry as "or", which will be interpreted as an operator in Connexion, and since it is likely to be at the end of a string, it will be an operator with bad syntax. More commonly, corporate or conference names may have words like "the" or "and," and personal names might have an "or" in uncertain dates, or AACR2 abbreviations that might not be recorded in the AR's variant (4xx) fields. Thus, some processing is still carried out.

Names-as-subjects are handled similarly, with the following query:

```
SELECT index_entry
FROM sierra_view.catmaint
WHERE condition_code_num=1 and field ~'^d6[0-3]'
```

```
;
```

Here the conditional selects fields beginning with a “d” (subjects) and the MARC tags 600 through 630. This therefore selects personal names (tag 600), conference and corporate names (tag 611 and 610), or uniform titles (630). In principle a MARC tag 620 could also be selected, but in practice this should not happen because 620 is undefined in MARC21.

And topicals are selected with a third query:

```
SELECT index_entry
FROM sierra_view.catmaint
WHERE condition_code_num=1 and field ~'^d65'
;
```

Here, any subject (d) tagged 65x is selected. UA’s implementation of Sierra tags only MARC fields 650 and 651 as subjects; 653 and 655 are placed in indexes with other tag codes.

The Beta method: processing

For each query, the output .csv files are opened in Editpad and the fourth line of regex from the Alpha process is used to remove stop words and operators. A minor hiccup was introduced to the process when batches of files processed by a vendor began to be loaded, as these included one or more subfield 0 in MARC 6xx fields. Because UA’s implementation of Sierra had not been set up to exclude subfield zero from indexes, the content of the subfield was included in the text. For example, a personal name subject access point for

Derrida, Jacques--Criticism and interpretation

uses the MARC coding:

```
600  10|aDerrida, Jacques|0http://id.loc.gov/authorities/names/n79092610|xCriticism and interpretation.|0http://id.loc.gov/authorities/subjects/sh99005576
```

and appeared in the index as:

> derrida jacques http id loc gov authorities names n79092610 criticism and interpretation http id loc gov authorities subjects sh99005576

An additional regex was needed to strip out the content of subfield zero: (http id loc gov authorities subjects sh[\d]+)|(http id loc gov authorities names n[a-z]?[\d]+). In the future, when the subfield zero is excluded from the indexes, it will not be necessary to remove these strings of characters. Thus, for this heading, after running the regexes to remove stop words and operators, and the subfield zero identifiers, the remaining data is:

> derrida jacques criticism interpretation

The text file is now ready for import into a Connexion batch search.

The Beta method removed a few steps from processing, and was also simpler in the sense that there was no need to cut vast selections of data from a single spreadsheet. This made the Beta method a bit less demanding of attention than the Alpha method.

Results

The Alpha method was tested on 91,491 entries in the headings report over a four month period (April 11, 2017-August 11, 2017). This ultimately yielded 31,891 ARs of all types. The Beta method was tested on slightly smaller number of entries -- 87,077 -- collected over a six month period (July 12, 2018-January 9, 2019). The results are summarized in table 3, Alpha and Beta results.

[place table 3 here]

The Alpha processing took a noticeably longer time to perform than the Beta, because the query results had to be sorted and saved into different files; the Beta process, involving just a

single regex substitution, could be performed rapidly. However the majority of the time needed for both versions was simply allowing the batch searching to run, exporting the ARs from Connexion, and loading the ARs into the ILS. Therefore the overall time spent on each method was nearly the same for a given heading report. The difference was more qualitative, as the Alpha method involved more attention to detail in selecting, reformatting, cutting and pasting, and saving data from spreadsheets. Notably, the Alpha query results always included some "junk" headings: non-MARC headings from brief bib records and headings in local 970 tags. The non-MARC headings were added to brief records by staff outside of the cataloging unit and were often incomplete; as they were not intended to be authorized access points it made no sense to search for matching authorities. The 970 tags had been added to provide access points for the table of contents of monographs and were in an idiosyncratic format which Sierra's automated authority processing (AAP) could not access. These "junk" headings had to be excluded from the batch processing as well.

To compare success rates, the number of successful AR retrievals is divided by the total number of entries searched in the batch to arrive at a success ratio. Comparing the ratio of success for the Alpha and Beta processes, the difference is rather small overall -- about a 35% success rate in the Alpha and 33% in the Beta. Differences emerge when comparing the success ratios for specific types of headings, and the total number of headings of each type. The Alpha data shows a 63% success rate for names, versus 49% in the Beta. The rates for names-as-subjects are closer, and based on smaller sample sizes. The rates for subjects are very small, at 5% and 9% respectively. One would expect less success in subject (and name-as-subject) searches because it is not often the case that extended strings of headings and subdivisions will match an identical and unique AR. Some ILSs will ignore subdivisions when verifying subject

headings, but UA's installation of Sierra checks the entire string including subdivisions. Similarly, name/title headings can pose problems, because NACO practice is not to create an AR for every title, but only those needing qualifiers or cross references. The issue is that the indexed fields lack subfield delimiters which would allow subdivisions to be removed before searching. While some subject authority records (SARs) are established with subdivisions, these are a minority of all SARs and the possible combinations of headings and subdivisions in bib records is vast. In principle one might search the batch of names-as-subjects in the subject index (su:). This would double the time and effort spent searching for names used as subjects, but it may be an avenue worth pursuing in the future.

The most glaring difference -- the difference in success rates for name entries -- may be explained by several factors.

First, the Beta process does not provide an easy way to remove AACR2 abbreviations from dates used to qualify names, such as "b." (for born) and "d." (for died). Because these would generally occur after a subfield d in the MARC field, the second regex in the Alpha could identify and remove them. But Beta selects indexed entries rather than the full MARC, so "b." and "d." in the names could be AACR2 abbreviations or they could be initials. It may prove helpful to devise a regex that will remove such abbreviations when occurring near numbers as a workaround.

Secondly, the Alpha and Beta test were not undertaken simultaneously. Because the Beta test was run later, it would likely be checking headings that do not have corresponding ARs. The entries in the later "Headings used for the first time" report would be less likely to have corresponding ARs simply because they are already being compared to a more robust authority

file in the ILS due to the ARs already loaded from the Alpha method. Ideally, the two methods should be compared using the same day's headings report.

Thirdly, there is the simple fact that the bib records loaded during the two test periods were different. This would be impossible to completely account for in principle, as different staff and faculty were loading different sets of bib records for different purposes in the normal course of the library's operation.

The higher success rate for name headings in the Alpha method is a problem requiring more investigation to explain. All in all there were far too many variables in the MARC ecosystem of a functioning ILS to make a truly controlled comparison.

Another complicating factor is that the second set of entries, which were used to test the second version of the process, had relatively fewer subject entries overall. This accounts for the similar "overall" success rates (35 and 33%) despite the Alpha process seeing significantly more successful name searches. This increases the suspicion that the difference in success rates owes more to the different bib records loaded than about the processes themselves. Thus, it was clear that a direct comparison of the methods was in order.

Alpha and Beta head-to-head

A more direct and meaningful comparison would be to run the two processes against a single headings report and compare the results. This comparison was made by allowing the "Headings used for the first time" report to accumulate for several weeks until there were 21,488 entries. Then both the Alpha and Beta methods were tested, with a stopwatch running to determine the exact amount of time the queries and processing took, beginning with opening the pgAdmin tool and stopping when the three files (names, names-as-subjects, and subjects) were saved. The results confirmed that the Beta method was considerably faster. The Alpha method took fourteen

minutes and two seconds. The Beta method took five minutes and forty-nine seconds. So, the Beta process clearly has the advantage in terms of time and effort.

In the single headings report, the Alpha and Beta queries yielded similar but slightly different counts for the total number of entries in each category. These are summarized in table 4, Search strings retrieved by the Alpha and Beta queries.

[place table 4 here]

The totals were reasonably close, but were not exactly the same. This discrepancy could be explained by two factors. First, some non-MARC entries from brief bib records made their way into the Alpha list. These still begin with an index tag of "a" so the Alpha query selects them along with the MARC fields. The non-MARC fields were obvious in the Alpha results, and omitted during the sorting. This would necessarily leave the Alpha lists shorter. But another unavoidable factor was that fourteen minutes had elapsed between the Alpha and Beta queries, so in effect the Beta query was querying a slightly larger report. Checking the report again after these tests revealed that another 40 entries had been added to the "Headings used for the first time" report since running the Alpha query. This small difference in total hits is tolerated as insignificant.

Running the two batches of results in Connexion yielded very similar results. The Alpha process had 4345 successful searches, while the Beta had 4348. De-duplicating the results of each batch reduced the hits for each to 4338 and 4341, respectively. Moreover, comparing the two sets to each other showed that the Alpha batch had 36 ARs not in the Beta batch, and there were 39 ARs in the Beta but not in the Alpha. Examination of the two sets of ARs did reveal some patterns to the discrepancies. These fell into two classes: conference name authorities and name/title authorities.

Conference headings were particularly problematic for both methods. Alpha returned the AR n50062132 (International Wheat Genetics Symposium) but the Beta did not. This can be accounted for by the fact that the MARC field which triggered the entry in the report was:

> a1112 |aInternational Wheat Genetics Symposium|0http://id.loc.gov/authorities/names/n50062132|n(12th :|d2013 :|cYokohama-shi, Japan)

Note that there is a subfield zero embedded in the heading. This is an artifact of an authority vendor's processing of the record for consortium that provides the e-resource bib reocrds. The expression **(\|e.*\|\|4.*\|\|0.*\|\|j.*)**, which was used to trim relator terms and URIs from fields, removed everything following the subfield zero. Thus the Alpha method batch searched only the portion in subfield a. Meanwhile the Beta batch searched the entire conference heading, including the specific numbering, year, and place. Because this was not established separately, there was no matching authority to return. A case might be made for wanting to retrieve the general conference name AR, even if it does not match a specific index entry, much as one might retain a topical AR for subjects that only appear further subdivided in the indexes.

On the other hand, the Beta method was able to retrieve the conference authority n 86042368, (Palestine Arab Congress. Executive Committee), while Alpha did not. In this case, the conference name uses subfield e for the subordinate unit (Executive Committee). In most 1xx and 7xx tags, subfield e is used for relator terms and therefore removed from the fields in the Alpha process. But because it used for part of the name in 111 and 711 fields, removing subfield e creates a less specific search, and the batch process, which accepts only single matches, rejected the multiple matches in the authority file. In this case, the unmodified Palestine Arab Congress and a specific meeting in 1921 were also established, making the Alpha version of the

search find three matches. Of course since only single hits are retained, none of these were retrieved.

It should be possible to further improve the queries to retrieve conference headings separately, so that they can be processed differently with revised regex substitutions and searched separately. This may be a project for the future.

In the case of name/title entries, there is a difference in how Sierra indexes name/title combinations and how they appear in the MARC fields. The MARC fields may be a single line, as in the case of 7xx fields with names in subfield a (and possibly qualifiers in b, c, d, and q) and titles in subfield t (and possibly qualifiers in l, s, etc.), or they may appear in two fields (1xx + 240). But Sierra only reports on names when a new name/title 7xx is added to the catalog. Therefore, the index entry retrieved in the Beta SQL query might be:

furman james 1937 1989

while the MARC field retrieved by the Alpha SQL query on the same entry is:

b70012|aFurman, James,|d1937-1989.|tHehlehlooyuh.

When the Alpha batch searches for "Furman James 1937 1989 Hehlehlooyuh," it will not find a match, but the Beta batch search for just the name will.

It would be possible to adjust the substitution regex to remove subfield t (and anything following it) for the Alpha processing, but this would be a two-edged sword: it will avoid missing some retrievable names, but it will also be unable to retrieve name/titles. This is ultimately a special case of the recall versus precision problem. Precision was favored in this case, and subfield t retained.

At this point it would seem that two processes are quite comparable in effectiveness. They have different strengths. Alpha can be a bit more precise. Beta is a bit less time-consuming.

Which is more suitable for use at a given library will depend on the resources that can be devoted to implementing and possibly improving them.

As a final test, the ARs from each set were loaded in "test" mode, so that a count of overlays (that is, ARs which are already in the catalog) and inserts (ARs new to the catalog) were reported without actually loading the ARs. The Alpha file had 746 overlays and 3592 inserts. The Beta had 749 overlays and 3592 inserts. Overlays would generally be "harmless" in the sense that at worst, they duplicated records already in Sierra. They might beneficially update an existing record, if the iteration in the catalog was out of date (pre-RDA forms, open dates, or changes to names). But inserts are generally the goal for this process.

One possible refinement would be to compare the relevance of the search results in terms of how many "blind references" the different methods produce. It is obviously likely that some of the successful batch searches will be "false hits" -- matches that are only partial and/or refer to different names or topics than the heading in use. The ARs thus retrieved will become "blind references," authorities that do not support any headings in bib records. Such blind references are normally suppressed or deleted in regular database maintenance. Anticipating this problem, at UA a MARC 910 field is inserted into all the ARs retrieved by the batch searches to identify the records as batch-loaded rather than manually added. This allowed us to select blind references in the "Blind references" headings report that originated from these batches for summary deletion.

The Alpha and Beta processes can help Sierra libraries, but is this in-sourcing approach applicable to other ILSs? The answer is yes, providing the ILS has some mechanism for reporting unauthorized or uncontrolled headings.

Gamma method for SirsiDynix Symphony

Another opportunity suggested itself with the SirsiDynix Symphony ILS. Symphony has a report that will export a text file identifying “unauthorized headings”. These are, like the headings in the Sierra report already discussed, headings that do not match any ARs in the system. Because the type of headings (names, topical, etc.) and even the MARC tags involved (100, 700, etc.) can be preselected before running the report, no SQL query or sorting is required. Moreover, name ARs can control both name access points and subject access points, so there is no need to search and load names-as-subjects separately from names. Figure 3 below is an illustration of a part of one such report’s output. Note that the report was run with the options to “format report” and “view log” unchecked -- leaving these options checked produces a less useable report with line breaks and page breaks that complicate normalization.

[place figure 3 here]

In the above sample, a few differences in the output from the SQL query will be evident.

First, there is the presence of some header information in the first few lines. These are generated by the system, and can simply be selected and deleted manually in the text editor. Secondly, diacritics in this report are displayed as an additional character, typically a character with a diacritic of its own. For example

Abū Dā’ūd Sulaymān ibn al-Ashʿath al-Sijistānī

is displayed here as:

Abåu Dåa®åud Sulaymåan ibn al-Ash°ath al-Sijiståanåi

The additional characters precede the characters that should have diacritics applied. This is a character encoding issue, as the Unicode encoding does not translate correctly into the output.

Third, a subfield “?” with the term “UNAUTHORIZED” is appended to each line. These appear in the staff view of bib records in Symphony as well. Finally, each line is preceded by a

number, indicating the number of occurrences of the heading in the database. Because these counts are before the MARC tags, it is important that the report be run for each type of heading (1xx/600/630/7xx names and 65x topical headings). But because Symphony allows NARs to authorize both name and subject uses of the name, there is no need to segregate names-as-subjects from names.

The other issues require some simple alterations to the regex used to normalize the data in the first Alpha processing. The first expression will remove the "counts" along with everything else preceding subfield a, and is fine as it is. Changing the third expression to **(\|e.*|\|4.*|\|0.*|\|\?.*)** will remove the "|?UNAUTHORIZED" along with relator terms, URIs, and the like.

Removing the characters representing diacritics is a bit more complicated, but is doable. After stripping out all the other MARC coding, the expression **[^a-zA-Z0-9- \x00-\x1F\x7F]** will select all the special characters standing in for diacritics. (The expression matches characters that are NOT letters, numbers, a dash, a space, or special characters like line breaks.) These are to be replaced with nothing (i.e., not a blank space).

The rest of the processing should be relatively obvious. Because the subfield delimiter symbol in both Sierra and Symphony is a "pipe," the other regexes in the Alpha processing will work the same way. Figure 4 shows the same entries of the sample report after running the following regex find/replace substitutions. The first and last expression should be replaced with nothing rather than a blank space.

1. (.*\|a)
2. (\|db\. ca\. |\|db\. |\|d\. ca\.|\|dd\. |\|dca\. |-ca\. |\|dfl\. ca\. |\|dfl\.)
3. (\|e.*|\|4.*|\|0.*|\|j.*|\|\?.*)
4. (\|.|\|$)

5. ("|;|:|\(|\)|\?| and | or |&c\.|&| in | an |,| the | for | on | so | with | to | by |”|’| be | that |\.{3}| near | same |\.)
6. [^a-zA-Z0-9- \x00-\x1F\x7F]

It would be reasonable to expect success rates similar to the Alpha method for Sierra, though this was not tested.

Future directions & further study

The initial success using the headings reports and batch processes to add ARs to the catalog has been encouraging. As the method has been tested by other librarians, additional tweaks and refinements to the query have been suggested. For example, David Green (Infrastructure Specialist, The State Library of Ohio) suggested working deduplication, processing, and removal of excess white spaces into the query by changing the three Beta queries to use the following first line[2]:

> SELECT DISTINCT trim(regexp_replace(index_entry, '(“|;|:|\(|\)|\?| and | or |&c\.|&| in | an |,| the | for | on | so | with | to | by |”|’| be | that |\.{3}| near | same | \s+)', ' ', 'g'))

Only unique (“DISTINCT”) entries are selected, and the regex substitution (with an additional term to replace multiple consecutive blank spaces) is applied to the output. A similar query, or set of queries, might be devised to speed up the Alpha method. Indeed this train of thought further suggests moving all processing out of the text editor program and into a batch of command line commands to further streamline the Gamma method as well -- an exercise perhaps for those with more advanced scripting skills than the present author. The ongoing development

[2] D. Green (personal communication, May 23, 2019)

of these methods through collaboration among librarians in different functional areas and different institutions has been gratifying and promises to further refine these methods.

Looking ahead, further manipulations such as removing subdivisions from subject entries should improve success rates. The detailed results log from the batch searches may also be worth examination to identify headings that should be checked manually when time or staffing permits. Moreover there is likely more that can be accomplished with other headings reports in Sierra (and other ILSs). For example, the Sierra "Near matches" report, which identifies entries that are partial matches to ARs, could be used to identify ARs that may need to be checked against the authority file (either the Library of Congress Name Authority File or OCLC) for updates. It may also be practical to use a SQL query to extract the "Correct heading is:" entry from the "Invalid headings" report, which notes fields in bib records that match variant forms in ARs. Batch searching the "Correct heading is:" entries would be a way to confirm that the ARs in the catalog are current, and re-loading them would trigger Sierra's AAP (at least in those Sierra implementations that have this feature turned on).

Further study is also warranted to determine the relative effectiveness of this method versus those achieved by different vendors, in terms of the number of headings correctly matched to ARs.

References

Aschman, A. (2002). The lowdown on automated vendor supplied authority control. *Technical Services Quarterly, 20*(3), 33-44. DOI: 10.1300/J124v20n03_03

Beisler, A., & Kurt, L. (2012). E-book workflow from inquiry to access: Facing the challenges to implementing e-book access at the University of Nevada, Reno. *Collaborative Librarianship, 4*(3), 96–116.

Carrasco, R. C., Serrano, A., & Castillo-Buergo, R. (2016). A parser for authority control of author names in bibliographic records. *Information Processing and Management*, *52*(5), 753–764. DOI: 10.1016/j.ipm.2016.02.002

Cook, D. (2014). Metadata management on a budget. *Feliciter*, *60*(2), 24–29.

David, R. H., & Thomas, D. (2015). Assessing metadata and controlling quality in scholarly ebooks. *Cataloging and Classification Quarterly*, *53*(7), 801–824. DOI: 10.1080/01639374.2015.1018397

Dong, E., Glerum, M. A., & Fenichel, E. (2017). Using automation and batch processing to remediate duplicate series data in a shared bibliographic catalog. *Library Resources & Technical Services*, *61*(3), 143–161. Retrieved from: https://journals.ala.org/index.php/lrts/article/view/6395/8442

Finn, M. (2009). Batch-load authority control cleanup using MarcEdit and LTI. *Technical Services Quarterly*, *26*(1), 44–50. DOI: 10.1080/07317130802225605

Flynn, E. A., & Kilkenny, E. (2017). Cataloging from the center: Improving e-book cataloging on a consortial level. *Cataloging and Classification Quarterly*, *55*(7–8), 630–643. DOI: 10.1080/01639374.2017.1358787

Heinrich, H. (2008). Navigating the currents of vendor-supplied cataloging. *IFLA Conference Proceedings*, 1–18.

Jackson, R. V. (2003). Authority control is alive and...well? *OLA Quarterly*, *9*(1), 9-12. DOI: 10.7710/1093-7374.1636

Kreyche, M., Lisius, P. H., & Park, A. (2010). The DeathFlip project: Automating death date revisions to name headings in bibliographic records. *Cataloging & Classification Quarterly*, *48*(8), 684–695. DOI: 10.0.4.56/01639374.2010.497721

Mak, L. (2013). Coping with the storm: Automating name authority record updates and bibliographic file maintenanc. *OCLC Systems & Services*, *29*(4), 235–245. DOI: 10.1108/OCLC-02-2013-0006

Martin, K. E., & Mundle, K. (2010). Cataloging e-books and vendor records: A case study at the

University of Illinois at Chicago. *Library Resources & Technical Services 54*(4), 227-237. DOI: 10.5860/lrts.54n4.227

Panchyshyn, R. S. (2013). Asking the right questions: An e-resource checklist for documenting cataloging decisions for batch cataloging projects. *Technical Services Quarterly, 30*(1), 15-37. DOI: 10.1080/07317131.2013.735951

Park, A. L. (1992). Automated authority control: making the transition. *Special Libraries*, *83*(2), 75–85.

Sanchez, E., Fatout, L., Howser, A., & Vance, C. (2006). Cleanup of NetLibrary cataloging records: A methodical front-end process. *Technical Services Quarterly*, *23*(4), 51–71. DOI: 10.0.5.20/J124v23n04-04

Snow, K. (2017) Defining, assessing, and rethinking quality cataloging, *Cataloging & Classification Quarterly, 55*:7-8, 438-455, DOI: 10.1080/01639374.2017.1350774

Tingle, N., & Teeter, K. (2018). Browsing the intangible: Does visibility lead to increased use? *Technical Services Quarterly*, *35*(2), 164–174. DOI: 10.1080/07317131.2018.1422884

Thompson, K., & Traill, S. (2017). Leveraging Python to improve ebook metadata selection, ingest, and management. *Code4LibLib*, (38), 1–17.

Tsui, S. L., & Hinders, C. F. (1999). Cost-effectiveness and benefits of outsourcing authority control. *Cataloging & Classification Quarterly*, *26*(4), 43–61. DOI: 10.1300/J104v26n04_04

Van Kleeck, D., Nakano, H., Langford, G., Shelton, T., Lundgren, J., & O'Dell, A. J. (2017). Managing bibliographic data quality for electronic resources. *Cataloging and Classification Quarterly, 55*(7/8), 560–577. DOI: 10.1080/01639374.2017.1350777

Vellucci, S. L. (2004). Commercial services for providing authority control: Outsourcing the process. *Cataloging & Classification Quarterly*, *39*(1/2), 443–456.

Williams, H. (2010). Cleaning up the catalogue. *Library & Information Update*, (Jan/Feb), 46–48.

Wolf, S. (2019). Automating the authority control process. Presented at the Ohio Valley Group of Technical Services Librarians Annual Conference 2019. Retrieved from https://uknowledge.uky.edu/ovgtsl2019/conf/schedule/17/

Wu, A., & Mitchell, A. M. (2010). Mass management of e-book catalog records: Approaches, challenges, and solutions. *Library Resources & Technical Services*, *54*(3), 164–174. DOI: 10.5860/lrts.54n3.164

Zhu, L., & von Seggern, M. (2005). Vendor-supplied authority control: Some realistic

expectations. *Technical Services Quarterly*, *23*(2), 49–65. DOI: 10.0.5.20/J124v23n02.04

Ziso, Y., LeVan, R., & Morgan, E. L. (2010). Querying OCLC Web Services for name, subject, and ISBN. *Code4Lib*, (9), 1–8.

Table 1. Summary of MarcEdit Validate Headings on selected record sets

Record source	**Bibliographic records loaded**	**Number of headings corrected**	**Corrections per title**	**% 1xx and 7xx corrected**	**% 6xx corrected**
EBSCO*	158,106	1,108	.007008	.002414	.000467
OCLC WCM	**209,777**	**5468**	**.026066**	**.008498**	**.000903**
Kanopy***	461,562**	20,447	.044300	.01123	.01346
Films on Demand***	10,064	1,273	.126490	.056359	.000162
Alexander Street Press	5,304	394	.074284	.012495	.008153

*EBSCO discovery layer records. These were often brief records with few or no access points, accounting for the relatively small number of corrections.

**Kanopy records were routinely re-loaded as a collection, at the vendor's recommendation, as corrections or changes to records were continuous. UA's set of Kanopy records was less than 20,000 titles, but the set was reloaded in its entirety monthly.

***Kanopy and Films on Demand records were also pre-edited with MarcEdit tasks that addressed certain recurring errors as mentioned above in the text. This somewhat decreased the overall number of changes made by the Validate Headings report, but nonetheless the rates of corrections are still greater than the OCLC benchmarks.

Table 2. Headings types in Sierra and WorldCat

Tagging prefix	Type of authority	Sierra index	WorldCat authority index	Sierra load table
a100	Personal name	Author	LC Names	Name authority
a110	Corporate body	Author	LC Names	Name authority
a111	Conference name	Author	LC Names	Name authority
b700	Personal name	Other author	LC Names	Name authority
b710	Corporate body	Other author	LC Names	Name authority
b711	Conference name	Other author	LC Names	Name authority
d600	Personal name	Subject	LC Names	Subject authority
d610	Corporate body	Subject	LC Names	Subject authority
d611	Conference name	Subject	LC Names	Subject authority
d630	Uniform title	Subject	LC Names	Subject authority
d650	Subject	Subject	LCSH	Subject authority
d651	Geographic name	Subject	LCSH	Subject authority

Table 3. Alpha and Beta results

	Names	Names-as-subjects	Subjects	Total
Alpha query	44410	5774	41307	91491
Alpha ARs retrieved	27935	1753	2203	31891
Alpha success rate (entries/ARs)	.629025	.303602	.053332	.34857
Beta query	49570	5916	31591	87077
Beta ARs retrieved	24424	1657	2893	28974
Beta success rate (entries/ARs)	.492717	.280088	.091577	.33274

Table 4. Search strings retrieved by the Alpha and Beta queries

	Alpha	Beta
Names	10771	10788
Names-As-Subjects	1403	1408
Subjects	9320	9339

Figure 1.

15	☐	**Headings used for the first time**	**Field:** b7001 |aAdolph, Wolfram,|earranger of music,|einstrumentalist **Indexed as AUTHOR:** *adolph wolfram* **Preceded by "a":** *adolph vincent r* **Followed by "a":** *adolphe bruce* **From:** b6097185x Bach, Johann Sebastian, 1685-1750, composer Rèveries

Figure 2.

a1001 |aChough, Sung Kwun,|d1985-	Chough Sung Kwun 1985-
a1001 |aChua, Hui Tong,|eauthor	Chua Hui Tong
a1001 |aChubb, Kit,|d1936-	Chubb Kit 1936-
a1001 |aCohen, Louis H.|q(Louis Harold),|d1906-|eauthor	Cohen Louis H Louis Harold 1906-
a1001 |aColombo, Maria,|eauthor	Colombo Maria
a1001 |aCranburne, Charles,|d-1696,|edefendant	Cranburne Charles -1696
a1001 |aCrozier, C. W.,|d1807?-|eauthor	Crozier C W 1807 -

Figure 3.

```
.folddata
.report
.report
.title
$(14810)

.end
.subtitle
$(14180)Wed May  1 10:02:02 2019

.end
.footing r


.end
    1  100: 0  : |a'ãAolåi,|eauthor.|?UNAUTHORIZED
    1  100: 0  : |aA mi.|?UNAUTHORIZED
    1  100: 0  : |aAQ,|eauthor.|?UNAUTHORIZED
    1  100: 0  : |aAbraham bar Hiyya Savasorda,|dapproximately 1065-approximately 1136.|?UNAUTHORIZED
    1  100: 0  : |aAbram,|cder Tate,|d1874-1962.|?UNAUTHORIZED
    1  100: 0  : |aAbåu Dåa®åud Sulaymåan ibn al-Ash°ath al-Sijiståanåi,|d817 or 818-889.|?UNAUTHORIZED
    1  100: 0  : |aAbåu Nuwåas,|dapproximately 756-approximately 810.|?UNAUTHORIZED
    2  100: 0  : |aAbåu al-Faraj al-Iòsbahåanåi,|d897 or 898-967.|?UNAUTHORIZED
    1  100: 0  : |aAbåu °Ubayd al-Qåasim ibn Sallåam,|dapproximately 773-approximately 837,|eauthor.|?UNAUTHORIZED
    2  100: 0  : |aAbåu °Ubayd al-Qåasim ibn Sallåam,|dapproximately 773-approximately 837.|?UNAUTHORIZED
    1  100: 0  : |aAce Hood,|d1988-|4prf|?UNAUTHORIZED
    1  100: 0  : |aAce Hood.|4prf|?UNAUTHORIZED
    1  100: 0  : |aAce Hood.|?UNAUTHORIZED
    1  100: 0  : |aAce.|?UNAUTHORIZED
    1  100: 0  : |aAchad,|cFrater,|d1886-|?UNAUTHORIZED
    1  100: 0  : |aAcharya Shunya,|eauthor.|?UNAUTHORIZED
    1  100: 0  : |aAchdâe,|d1961-|?UNAUTHORIZED
    1  100: 0  : |aAding,|d1972-|eauthor.|?UNAUTHORIZED
```

Figure 4.

Aoli
A mi
AQ
Abraham bar Hiyya Savasorda approximately 1065-approximately 1136
Abram der Tate 1874-1962
Abu Daud Sulayman ibn al-Ashth al-Sijistani 817 818-889
Abu Nuwas approximately 756-approximately 810
Abu al-Faraj al-Isbahani 897 898-967
Abu Ubayd al-Qasim ibn Sallam approximately 773-approximately 837
Abu Ubayd al-Qasim ibn Sallam approximately 773-approximately 837
Ace Hood 1988-
Ace Hood
Ace Hood
Ace
Achad Frater 1886-
Acharya Shunya
Achde 1961-
Ading 1972-

List of figure captions