ARL Statistics Best Practices for Deduplicating Serial Titles
December 4, 2008

Deduplicating through Sampling (see also Q5 in Statistics FAQ for the following language):

If it is not possible to deduplicate across libraries due to system limitations, a random sample of 1,000 serial titles can be generated from the branch library and an analysis of the overlap conducted. The resulting percentage overlap can be extrapolated to all the number of titles available for that branch to get an estimated figure of the title overlap for all serials held by a specific branch. Subtract the estimated title overlap from the main serial file.

Strategies for Deduplication

Attached are 4 documents that can be used as examples for deduplicating serial title counts.

- Using the SFX Knowledgebase to obtain ARL E-Journal Count
- Best Practices: Using SerialsSolutions (SS) Knowledgebase for De-Duping/Counting Electronic Serials
- Library catalog as the main source of records
    - Programming capability (Columbia University)
    - University of Chicago

Using the SFX Knowledgebase to obtain ARL E-Journal Count

As with the SerialsSolutions best practices, use of SFX makes the basic assumption that "All of your institution's electronic serials, both purchased and not purchased, are being tracked within the… knowledgebase," and Government Documents numbers are counted separately. Print serials numbers need to be separately de-duplicated so that only unique titles in print are counted. Serial 'databases' are not included in this count. Libraries with holdings greater than 65K titles before de-duplication will want to use Excel 2007, which allows over 1 million rows.

Total Unique Title Count for E-Journals

1. Use SFX tools to export a text file containing all active 'object portfolios'.

   [Export instructions: From SFX Admin Center, choose KnowledgeBase Tools -> Export Tool -> Advanced Export Queries.
   Select following options for export: Output Format = TXT, Export Object Type = Serials, Export active portfolios with the following services = getFullTxt, Export from ALL targets.]

2. Open resulting text file in Excel and remove all columns except for Title and ISSN. Split list into two separate worksheets, one worksheet for titles with an ISSN and one worksheet for titles without an ISSN.

3. Remove Title column from list of titles with ISSN. Sort file by ISSN column and then filter by unique records only. Copy resulting list and paste into new worksheet. This will give total number of unique titles with an ISSN. [29,257 unique titles]

4. Remove all but Title column from list of titles without ISSN. Sort file by title and then filter by unique records only. Copy resulting list and paste into new worksheet. This will give total number of unique titles without an ISSN. [13,407 unique titles]

5. The *Total Unique E-Journal Title Count* is the sum of unique (de-duplicated) e-journal titles with ISSN and without ISSN. [42,664 unique titles]

   [Note: This number could be used as a base for Question 4, but may not be needed if the Library intends to answer Q4a and 4b instead. Law/medical titles would need to be de-duplicated from this number, and numbers added for print-only unique titles and serial database subscriptions.]

Unique 'Currently Purchased' Title Count for E-journals

1. Use 'Export Tool' from SFX Admin Center in order to export text file containing all active object portfolios from 'subscribed' targets.

   [Export instructions: From SFX Admin Center, choose KnowledgeBase Tools -> Export Tool -> Advanced Export Queries.
   Select following options for export
   Output Format = TXT
   Export Object Type = Serials
   Export active portfolios with the following services = getFullTxt,
   Select only subscribed serial targets for export.
   Inclusion/Exclusion for Notre Dame is as follows:
   Include all paid fulltext aggregator targets and all paid publisher targets. Exclude all targets for resources received through state-funded Inspire databases (Ebsco full text databases, Informe). Exclude 'targets' which comprise proceedings, whether using issn or isbn (IEEE, ACM)

2. Open exported text file in Excel. The *Total Unique Subscribed Title Count* will be the sum of unique titles with ISSN and unique titles without ISSN.
   - In Excel, remove all columns except for Title and ISSN.
   - Split list into two separate worksheets, one worksheet for titles with an ISSN and one worksheet for titles without an ISSN.
   - Remove Title column from list of titles with ISSN. Sort file by ISSN column and then filter by unique records only. Copy resulting list and paste into new worksheet. This will give total number of subscribed (purchased) titles with an ISSN.
   - Keep only Title column from list of titles without ISSN. Sort file by title and then filter by unique records only. Copy resulting list and paste into new worksheet. This will give total number of purchased titles without an ISSN.

3. Law De-duplication: Obtain ISSN and non-ISSN lists for Law titles. Instructions are the same as for the *Currently Purchased Title Count* except for the 'Export from SPECIFIC targets' options in the Advanced Export Queries. In this export we include only resources purchased by the Kresge Law Library (Hein Online and LexisNexis). Note: Law Library says this represents most of their subscribed journal titles, and they are unable to give us a more complete list.

   ISSN and non-ISSN Title lists from these targets are de-duplicated from the Currently Purchased E-Journal lists. It is not sufficient to simply exclude Law title 'targets' from the *Currently Purchased* list, because some Law titles are provided independently from other providers.

Save the two lists of currently purchased e-journal titles, with and without ISSN. They will be needed for addition of print-only titles, and for later de-duplication against the list of free titles.

Total Unique Free Title Count

1. Use 'Export Tool' from the SFX Admin Center to export a text file containing all active object portfolios from Free targets.

   Follow export instructions for *Unique Subscribed Title Count* but:

   - Include ONLY free/open access targets in the 'Export from SPECIFIC targets' option in the Advanced Export Queries.
   - Inclusion/Exclusion for Notre Dame is as follows. Include all active free/open access targets (DOAJ, BioOne) and all targets for full text databases acquired through Indiana state funded INSPIRE Project (EBSCOHost full text databases, Gale Informe).

2. Exported text file will be opened in Excel and two unique title lists (with and without ISSN) will be created.

3. De-duplicate lists against 'Subscribed' titles lists, excluding titles that are duplicated in the '*Subscribed Title Count*.'

   [Method: Use Excel *VLOOKUP* feature to identify all titles in the list of free titles with ISSN that are also on the list of subscribed titles with ISSN. The same VLOOKUP process should be used for the list of free titles without ISSN against the list of subscribed titles without ISSN. Remove all titles from the list of free titles that were found within the list of subscribed titles.]

SFXWorking Definitions

- Object Portfolio – a single instance of a journal title. A journal title which is included in several provider packages, such as *J of Acad Librarianship*, can have more than one object portfolios, including Elsevier, Ebsco, Gale, etc.

- Target – a logical group of titles identified by either a publisher (Elsevier), provider (Project Muse), platform (ScienceDirect, SpringerLink, EBSCO, Metapress, DOAJ) or type (IEEE Proceedings, ACM Proceedings). The power of SFX targets is that they can be customized to reflect institutional needs—they are not frozen and limited by global definition. A group of titles can have a locally created 'target' which is not included in the global knowledgebase distributed by ExLibris (eg. BNA law titles). It is possible to split a database such as EBSCO Academic Search Premier into

segments, so that one level reflects 'free' titles and an add-on segment might reflect 'purchased' titles. IEEE can include separate targets for journals and proceedings. Judicious use of targets can assist the process of putting particular groups of titles into different categories (targets) based on local criteria. Object portfolios for different instances of the same journal title can exist in more than one target (eg J of Acad Librarianship can be found, with varying coverage, in several targets).

ARL Report Requirements:

Q4a Serial Titles purchased:

Include unique, de-duplicated print/mic/electronic journal holdings. Exclude non-purchased titles provided in full-text databases by state funds, as well as titles that are considered Open Access or free to any user. Exclude titles paid for by Law Library.

Q4b Serial titles not purchased:

Include INSPIRE titles and free titles; in case of duplication, title is counted under Q4a, paid titles.

Include government document serials if possible (or note in Q5)

Print Titles:

Unique print serial and standing order title lists are exported from our ILS system. Fields include Order type; Order number; Material type; Order Group; Order status; Method of Acquisition; Title; Imprint; ISSN,ISBN. Relevant codes used include electronic in order type and print+electronic in method of acquisition. Excel is used to de-duplicate any remaining title or issn overlap.

Serial 'databases' are obtained by coding maintained in the ILS and exported as part of a special datamart process which allows databases coded as 'Material Type' (Reference or E-Text) and 'Order Type' (Serial) to be exported and analyzed for appropriate inclusion.

Best Practices: Using SerialsSolutions (SS) Knowledgebase

for De-Duping/Counting Electronic Serials

There are some basic assumptions that should be tested before going forward with this best practice: All of your institution's electronic serials, both purchased and not purchased, are being tracked within the SS knowledgebase; all of these titles are being served up to your users either from an A-Z list or from bibliographic records in your ILS; and "all" does not include Government Publications which are now being counted separately

⇒ Step 1. If you are tracking the Government Documents that SS includes in their Knowledgebase, be sure to deselect this so those titles are not included in your Data on Demand File. This will temporarily exclude these titles from your A-Z list if you are using this SS functionality. Another option is to leave this as selected, but once you get the Data on Demand File, remove the titles from the spreadsheet and paste them to their own file. You may be able to use that file later when working to compile serial Government Documents counts and this method will not disrupt user access to the titles via your A-Z list.

⇒ Step 2. Request a special Data on Demand Report from SerialsSolutions – request that SSJ ID and SSJIB numbers be included in the report. [SSJ ID = SerialsSolutions Journal Identification numbers; SSJIB = SerialsSolutions brief record control numbers]

All SerialsSolutions clients can request this report even if you do NOT subscribe to the SerialsSolutions MARC record service--this was confirmed with SS in September 2008. It takes a few weeks for SS to create and provide the report so think ahead and give yourself plenty of time to get the report.

⇒ Step 3. Make sure your Data on Demand Report is exported/saved in an Excel file, preferably 2007. Excel 2007 accepts bigger files and the de-duping function is easier than using the advanced filter function in Excel 2003.

⇒ Step 4. Sort the spreadsheet by "Provider" column as designated by SS. Sorting by Provider should facilitate the identification of titles as purchased or not purchased.

⇒ Step 5. Insert a new column into the spreadsheet. This new column can be labeled whatever you choose, something like Purchased/Not Purchased. Have the appropriate

person from your Library annotate the cell next to the title as purchased or not purchased. Do this in such a way that purchased will sort ahead of non-purchased (e.g. 1 = purchased, 2 = not purchased). An alternate mode of identifying Purchased/ Not Purchased titles is to look to see if your URLs contain a proxy code. Assume that those records containing URLs that route access through the proxy client are paid and those items providing direct URLs without proxy mediation are free. An acquisitions or serials librarian knowledgeable about payments should scan the results to test this assumption.

⇒ Step 6. Re-sort the spreadsheet by the Purchased/Not Purchased column so Purchased titles rise to the top of the spreadsheet and Not Purchased fall to the bottom. Do a secondary sort on ISSN so those with ISSNs are preserved over those not containing ISSNs. In the de-duping process, Excel will retain the first instance of a match over subsequent instances.

⇒ Step 7. De-Dupe the file using the SSJ ID & SSJIB numbers; using Excel 2007's "Remove Duplicates" function on the data ribbon, or Excel 2003's advanced filter function under the "Data" function.

⇒ Step 8. Resort this data by ISSN. Cut all records from this spreadsheet containing ISSNs and paste them on a second spreadsheet. Sort these records by Purchased/Not Purchased. De-Dupe the file using the ISSNs using Excel 2007's "Remove Duplicates" function on the data ribbon, or Excel 2003's advanced filter function under the "Data" function. Copy the records left after this operation back into the original spreadsheet.

⇒ Step 9. Resort the data now in the original spreadsheet by eISSN. Cut all records from this spreadsheet containing eISSNs and paste them on a second spreadsheet. Sort these records by Purchased/Not Purchased. De-Dupe the file using the eISSNs using Excel 2007's "Remove Duplicates" function on the data ribbon, or Excel 2003's advanced filter function under the "Data" function. Copy the records left after this operation back into the original spreadsheet.

The resulting file will contain a list of unique electronic titles, with purchased being privileged over not purchased.

⇒ Step 10. Re-sort the original spreadsheet by Purchased/Not Purchased column.

⇒ Step 11. If the library has electronic serial titles not tracked in the SerialsSolutions Knowledgebase but access is made available to users, (for example, serial databases without full-text), the library will have to determine how to add them to the resulting

count derived from the Excel spreadsheet- remember to annotate as Purchased/ Not Purchased.

The resulting file can be used to respond to Questions 4ai (Electronic purchased) and 4bi (Electronic not purchased) on the ARL Annual Survey form. Individual libraries will have to determine how to compile and de-dupe their Government Documents (electronic and print) and their print unique title counts from their ILS, SerialsSolutions files or other mechanisms.

One important caveat – DO NOT employ the SerialsSolutions Overlap Analysis function to derive a unique electronic title count. The overlap analysis count that is provided per this function is truly a count of "unique titles" tracked within your library's knowledgebase. If a title is included in more than one "provider" package it is eliminated from the count entirely and then not counted at all.

Document compiled by:

Paul Beavers, Wayne State University, aa6536@wayne.edu

Betsy Redman, Arizona State University, bjredman@asu.edu

Deborah Sanford, University of Connecticut, deborah.sanford@uconn.edu

10/06/2008

Columbia University Libraries

Serials Counts for 2006/2007

Columbia began by running an analysis of the information on serials we hold in CLIO (Voyager ILS). One of our Systems Analysts very kindly wrote up the process he used to de-duplicate the records in CLIO. (See below.)

Since there is minimal overlap in our collections, we simply added in the serial titles from the Law Library. After reviewing the information, from Teachers College, we reviewed their databases and only added the titles that we didn't receive in the Morningside operation reflected in Voyager.

De-duplication Process

Systems Analyst builds a title match key for each record. Each component of the key is sent to a normalization routine which:
-- converts special characters to their normal equivalents (e.g., Polish L to 'L')
-- removes non-filing characters
-- removes extra spaces
-- uppercases the string

The match key is built as follows:
1) subfields a, n and p of the 245 are normalized;
2) if there is a 100, 110 or 111, the normalized version is prepended to 1;
3) if there is a 130 " (Online)" is removed and it is normalized; if it differs from 1, the normalized version is prepended to 1.

The match keys are then used to group titles together.

The analyst ran two analyses, one based on title match, the other on ISSN match. The end results are nearly identical.

To refresh your memory about method:

Six characteristics were tested for each title group:
ordrs = active order
paymt = payment posted since 7/1/06
chkin = checkin made since 7/1/06
er ss = electronic resource, serialssolution record
er ot = electronic resource, other record source
items = item record created since 7/1/06

A title group was rejected if it represented a dead serial (based on 008 date type = 'd') without an electronic version. For the remainder, the title group was considered active if any of the six categories above was characteristic of the group.

University of Chicago

Working assumptions:

1. Include all "currently received" serials, regardless of format or whether or not they were purchased.

2. Include each title only once, regardless of multiple subscriptions or versions. (A title available in print, via publisher-based online access and also via JSTOR counts as ONE title).

3. Be comprehensive, and count any title for which we're provided access, either in our catalog or through another finding aid, such as SFX.

Our procedure:

Chicago is a highly-centralized library, and all branches share a single online catalog. Except for a statistically insignificant number of exceptions, all our holdings for a given title are on a single bibliographic record. The catalog contains a comprehensive record of all titles currently received in print as well as all electronic titles to which we individually subscribe. Our practice is to put the print and electronic version on a single bibliographic record eliminating duplication there. There are a very small number of titles that we actively receive in both print and microform editions and these are on separate records, but there are very few of these remaining.

Our practice is to code the bibliographic records for "active" serials in such a way that we can easily get an accurate count of our currently received serials. We don't need to rely on indicators such as an open date in the fixed field, or through some other convoluted combination of search criteria. We have a single distinct local code we can use to identify those titles that are currently being received.

Although many aggregated titles are represented in the catalog, they donot contain the activity code and are therefore excluded from the catalog count. These are counted in the next step, when we include SFX figures.

The current scan of the catalog reveals the unique titles that are currently being received - print and/or electronic. We are very confident in this number.

The second source we use is our SFX database. Many, but not all [of the entries] … are also in the catalog.

We take the list of unique titles from SFX and dedup that against our catalog records, using ISSN as the match point.

Currently, this yields titles that are accessed via SFX but that not in our catalog as "active". (Many of these actually DO have aggregator records in the catalog but aren't in the set we pulled from the catalog since they aren't coded as "active".)

We then add the two numbers (titles from the catalog plus titles from SFX) to arrive at a total current serial count.

In summary, our process is:

1. Count active titles (unique) in our catalog
2. Count active titles (unique) in our SFX database
3. Dedup the two sets by ISSN.