Do-It-Yourself Title Overlap Comparisons

Melissa Belvadi
University of Prince Edward Island, mbelvadi@upei.ca

Melissa Belvadi, "Do-It-Yourself Title Overlap Comparisons" (2015). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284316261

Do-It-Yourself Title Overlap Comparisons

Melissa Belvadi, Collections Librarian, University of Prince Edward Island

Abstract

Discovery service indexing content can be highly customizable, which makes traditional title overlap analysis published by third parties less meaningful for a library making new subscription or cancellation decisions. This article presents a method for conducting basic title overlap analyses in-house at minimal cost, tailored to the specific configuration of the library.

Why DIY Comparison?

Whether it be for new programs, new competing products, or cancellations, librarians make decisions about which abstracting and indexing (A&I) and full-text products to subscribe to. Often, we turn to our scholarly literature, where colleagues have performed painstaking comparisons between specific products. Readers could be confident that if they licensed the same product, they would have access to the content that was tested in the article.

However, we now have a new kind of finding tool: the discovery service (DS). DSs are used by many library patrons for research as they would previously have used a single A&I database. They are substantial investments, so librarians making budget-driven subscription decisions need to take their DS into account.

But DS content is highly customizable. This article uses EBSCO Discovery Service (EDS) as the exemplar, but other DSs may be similar.

A library EDS license starts with a basic index, which includes over 250 million citations.

The library then adds whatever databases, A&I or full-text, are licensed with EBSCO. That includes both EBSCO-produced and third-party products, like EBSCO's Business Source, CINAHL, and SocIndex, as well as Medline, ERIC, PsycINFO, and MLA.

Then the library can choose to add any from hundreds of other sources available from third parties. Some, like the Harvard Bibliographic Dataset, are free. Others, such as MathSciNet and Web of Science, are only an option if the library is a mutual customer.

Finally, the library can provide its own metadata to include, known in EDS as a "custom catalog." The Library's literal catalog is the usual first choice. But they can provide other sets such as the metadata from their institutional repository.

EBSCO reports that there are about 850 sources that can be added to the basic index, with up to about 2.5 billion records as of October 2015. It is likely, then, that no two EDS libraries have the same search content available, even aside from the custom catalogs.

So the traditional methods of comparing two products, by either searching sample topics or title overlap, do not provide useful data for other libraries if one of those products is a DS. Each library needs to conduct its own analysis against its DS configuration.

Why Title Overlap Analysis?

There are two common ways to compare A&I products: sample topic search results analysis (STSRA) and title overlap analysis (TOA). STSRA involves searching the products side-by-side with the same set of search topics and then comparing the results, usually by some combination of recall (looking for known good articles that should be in the results) and precision (quality of the relevance of the results). STSRA is the ideal, but to do it effectively requires both significant subject expertise and a substantial number of discipline-specific sample topics that are appropriate for the research demands of the institution. Perhaps a grant can be obtained on rare occasions, but that

http://dx.doi.org/ 10.5703/1288284316261

would not be a method for everyday use. So TOA is the next best option.

TOA involves comparing the indexing coverage of specific journal titles, taking into account depth of indexing (cover-to-cover or selective) and years of coverage.

What TOA does not do is compare the quality of the metadata available in each product. Thus it has serious limitations for collection decision-making. The quality of metadata, particularly controlled-vocabulary subject terms, is very important to a researcher's success. The power of a professionally maintained and carefully applied subject thesaurus should not be taken lightly in such decisions. Some databases, like CINAHL, PsycINFO, and Historical Abstracts, have specialty metadata (e.g., the age and gender of research subjects, the time period of the historical research, etc.) that is valuable to researchers in those specialties in producing high quality results. Because a DS is designed to search literally hundreds of millions or even billions of records, there has never been a time that surfacing to the top of the relevancy ranking the best possible results has been more important.

So TOA should be considered just one tool. In particular it can be helpful in the negative result case, when there is such insufficient overlap that the product under consideration would definitely provide unique indexing content. In that situation, the more difficult STSRA is unnecessary because in the absence of adequate content, significantly poorer results in terms of "recall" is a certainty.

What About Google Scholar?

The one DS that is available for free to all library patrons is Google Scholar (GS). Increasingly, faculty and student researchers are turning to GS before library-licensed databases, with one study suggesting that close to 20% of researchers are using GS first (Hightower & Caldwell, 2010). It is possible that such choices are far more widespread than many librarians realize. So librarians testing A&I product content may want to include GS in their list of "competitors" along with their DS.

Workflow

UPEI's Robertson Library staff have developed a TOA workflow to compare products using Google Sheets, which is very similar to Microsoft Excel, and undergraduate student assistants who are hired primarily to provide technical support to patrons on a scheduled-shift basis. This workflow has been used to evaluate several existing products as well as products under consideration for new subscriptions. This fills up the downtime that students at a service point have when not helping patrons and thus does not require any new funds. Depending on the time of semester and length of the title list, the 10-hour-per-week IT student assistants can complete a TOA in just a few weeks. The task is ideal for that work environment, because the task is done cell by cell, so there is no loss of accuracy when they pause and resume around patron interactions.

Selecting Titles to Compare

TOA does not always mean evaluating every title covered in an index. Every institution has its own research areas of interest, and in some cases, an index may include many journals that are not within that scope.

Further, if the analysis involves an existing subscription, the library may have fairly precise title-level usage data about citations or abstracts viewed by its users. The librarians may thus restrict the analysis to those titles with demonstrated interest by their own patrons.

Some index publishers divide their lists into "core" titles that are fully covered versus "selective" titles which only include specific articles relating to the theme of the index. It is reasonable then to consider only conducting the overlap analysis on the "core" titles, with perhaps a small random sampling of the selective titles.

Finally, when considering products to support an entirely new program area, the primary concern may be to make sure the index covers the "most important" journals, whether that be judged by the faculty or by various external metrics such as impact factor or H index.

This last method is especially useful because unlike the usual methods, which tend to be one-sided in starting with a specific product's content and comparing the DS and GS to that (a process that is inherently biased in favor of the product), this one allows one to discover weaknesses in the product as well as the comparison targets.

However the title list is compiled, they need to be "cleaned up" as much as possible with regard to abbreviations, subtitles, punctuation, and the like, especially if GS will be included as a comparison target, because GS does not provide ISSN searching.

Spreadsheet Use

The recommendation is to have one spreadsheet file with one "master" worksheet, which contains the list of titles of interest and where the analysis will take place. Other worksheets will contain whatever other lists may be used to compare data, such as the coverage data for the product, or title lists of other competing products if the analysis will involve more than just the DS and GS.

A general principle for working with spreadsheets is that when an entire column contains formula-generated values, and the data it is drawn from is not going to change, use "autofill" to populate the entire column, open up a new empty column next to it, copy the entire formula-generated column, then use "Paste Special – values only" to copy all of the data into the new column, finally deleting the original formula-generated column. The reason for this is that maintaining large numbers of formulas (each cell down the column is a separate formula) significantly impairs the response time of the spreadsheet. Tip: autofill only works if it is immediately to the right of a column that has data in every row (otherwise it stops at the first blank).

Handling ISSNs

After establishing the desired title list, the second step is to reconcile ISSNs among the products. Generally the best way to get accurate results is by ISSN searches. However, title lists and data sources can vary as to whether they use the print ISSN, electronic ISSN, or both. Also, the absence of the hyphen in the list can easily cause spreadsheets to treat the ISSNs as whole numbers and drop leading zeros. Here are some tips for getting ISSN columns useful for testing:

1. When importing csv/tsv data into Excel, do not open it directly with Excel. Instead, start with a blank spreadsheet and use Data – Import from file to insert the data into the worksheet. That allows you to specify that the ISSN column(s) be treated as "text."
2. In Google Sheets, if the ISSNs were converted to numbers upon upload, create a corrected column using a formula like this: =text(A2, "0000-0000") (then autofill the column down) where column A has the ISSNs listed as numbers that have lost their leading zeros.
3. If your ISSNs lack hyphens, open a new empty column and use the following formula to add them. This assumes your list starts in row 2 and the ISSN is in column A: =left(A2,4)&"-"&right(A2,4) . Normalizing all of the ISSNs to include the hyphen is recommended because it guarantees that the cell contains text, and not a numeric value.
4. If you have two columns of ISSNs, for print and electronic, normalize both.

Formulas for Data Entry

The third step is to set up the columns to be filled in. Here is an example of a set of columns to use when comparing a list of titles for a potential new index, to GS and to EDS:

- The start/end date coverage of the index being tested:
    - If the vendor provides a list to pull this data from, upload it as another worksheet:
        - Normalize the ISSN data as above.
        - Create a single column combining the start and end indexing dates. For instance, if the start date is in column F and the

end column is in column G, use: =F2&" - "&G2.

- Use either index/match or vlookup to insert this data into the column in the main sheet:
  - Vlookup is easiest if the ISSN column comes before the coverage range. For instance, if your vendor coverage worksheet is called “db” and the ISSN column you will match on is in column C in “db” and column B in the main worksheet, and the combined coverage range is column H, use this formula in the new column on the main worksheet: =vlookup(B2, db!C:H,6, 0) (then autofill). The “6” refers to the fact that column H is the sixth column in the range from C to H, and the “0” tells Excel or Google to only report back an exact match.
  - Index/match can be used instead of vlookup if the coverage column comes before the ISSN column. The equivalent index/match formula for the vlookup situation above is: =index(db!H:H,match(B2,db!C:C,0)). Note that since columns C and H are specified without counting relative position, this method would work if they were switched, whereas vlookup requires the ISSN column to be before the coverage column.

- The number of hits for the title in the index (if available, e.g., a current or trial subscription).
  - If it is possible to create a column of direct persistent links, by either ISSN (ideal) or journal title, into the index, do so as in the examples for GS and EDS below.
  - Otherwise try to give the student as complete a search text as possible to save time and error. For instance, with Proquest’s interface, which as of the time this was written did not allow the generation of persistent links by formula, we created this column, and had the student just copy and paste each into the Proquest search box (this example illustrates how to combine both the print and e-ISSN if you have both for maximal accuracy): ="ISSN("&B2&") OR ISSN("&C2&")".
  - The student can retrieve both the total hits and range of coverage years as separate data columns if coverage years were not available as per above.
- Link to search GS by journal name
  - Assuming the title is in column A, and the assistant gets the word “Search” to click on: =hyperlink("http://scholar.google.ca/scholar?as_vis=1&hl=en&as_publication="&A2&"&as_sdt=1,5","Search").

    - GS does not offer an ISSN search, so consider results of very short titles that may be substrings of longer journal titles carefully.
- Hit count in GS (result of using the "Search" links): GS allows for a quick filter of results by 2011+, 2014+, or 2015+, or a custom range or just the count of all of the results.
- Link to search EDS by ISSN (in the B column):
    - =hyperlink("http://search.ebscohost.com/login.aspx?direct=true&bquery=IS+("&B2&")&clv0=Y&type=1&site=eds-live&scope=site","Search").
    - If you have both print and e-ISSN, use the print column for EDS searches.
- Hit count in EDS—total results as it appears at the top, which is already deduped.
- Range of dates covered in EDS—the student can see at a glance the starting and ending dates of coverage. This could be important because sometimes the specialty indexes cover earlier years than the dataset that EDS is using does.
- If the index you are testing is actually a part of your DS, like Historical Abstracts is for UPEI's EDS, create a test profile within the DS, which includes everything that your live one does except the one index you are testing. In EDS this is very easy to do as the administrative account service makes it easy to copy an entire profile's settings and then just turn off that one database. In this situation, modify your EDS search link to refer to that profile instead of your default live one, by adding: "&profileid=" and the code you named it in the administrative service.
- If the product being tested is a full-text product, you may also want to include a column with your current full-text holdings. If you can provide a "Search" link column into your holdings (e.g., your A to Z service), provide that column and the separate column for the holdings summary itself.
- You may also have a column indicating if the indexing is "core" in the product being tested, or if it is important to your program, or peer-reviewed.

Once the spreadsheet is set up, complete the first few rows to provide examples to your student assistants. It is also a good way to troubleshoot any problems with linking formulas. Then give it to the assistants to work on.

Analyzing the Collected Data

The fourth step is to analyze the data. How to do this will depend on which columns you have used.

- The easiest first step is to use Data – Filters to see which rows have 0 hits in your DS or GS. If there are many, and if high importance titles are among them (which you also use Filter to see), this might just decide the matter without further analysis.
- You may also want to consider a very small number of hits to be effectively 0. There are a couple of ways to do this depending on if you are using Excel or Google Sheets. But the one that gives the most control is to create calculated columns with criteria for then filtering on.
    - For instance, if you want to see all rows where both GS and EDS have fewer than 100 hits, you could use this formula, where the GS hit count is in column G and the EDS hit count is in column I: =if(and(G2<100,I2<100),0,1). In this example, the column will show 0s when both GS and EDS are below 100, which you can then use Filter on to show matching rows.
    - You could also use COUNTIF at the bottom of that column to get a gross count of the zeros, e.g.

=countif(J2:J250,"0") if the above is in column J from row 2 to 250.

- If you have importance designations (e.g., if you labeled some titles “core”), you may want to compare the index’s hits to GS and EDS and count up only for the core titles; if column C has the ‘core’ label and column D has the index’s hits, and G and I as above: =if(and(C2="core",G2<D2,I2<D2),0,1).
- If you prefer more readable labels instead of using 0,1 to indicate criteria matches, you can use words like “good” and “bad” in place of the “0,1” at the end of the IF (use the quotation marks around them); countif will accept strings, e.g., =countif(J2:J250,"good").

Endnote: All spreadsheet formulas with more detailed instructions are available at: https://goo.gl/hgJoJi

References

Hightower, C., & Caldwell, C. (2010). Shifting sands: Science researchers on Google Scholar, Web of Science, and PubMed, with implications for library collections budgets. *Issues In Science & Technology Librarianship, 63*. http://dx.doi.org/10.5062/F4V40S4J