Striving for Uniqueness: Data-Driven Database Deselection

Jeremy M. Brown
Mercer University Libraries, brown_jm@mercer.edu

Geoffrey P. Timms
Mercer University

Jeremy M. Brown and Geoffrey P. Timms, "Striving for Uniqueness: Data-Driven Database Deselection" (2012). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284315104

Striving for Uniqueness: Data-Driven Database Deselection

Jeremy M. Brown, Head of Library Systems, Mercer University
Geoffrey P. Timms, Electronic Resources and Web Services Librarian, Mercer University

Introduction

As libraries endure an ongoing crisis of available funds to meet inflating electronic content costs, the hatchet is kept ever close at hand to dispatch the perceived least important e-resources to help balance the budget. One school of thought is to eliminate index/abstract databases to preserve full-text periodical content. Another is to continue to maintain a balance between discovery and access. At Mercer University Libraries, we recognize this now familiar challenge of finding areas in which to trim the fat. We are forced to look ever closer at our subscriptions to prioritize our patrons' needs, maintain budgetary equilibrium, and remain true to our goals; yet we've already eliminated the easy targets. The Library Systems Department has worked to develop a tool to assist decision makers with pertinent information about the uniqueness of both our full text and index databases and packages.

When difficult decisions must be made, especially those which will alter the ability of an academic program to conduct research, it is important to provide data which supports the conclusion. Data helps to dispel myths or sentiments which cloud the honest evaluation of a resource. Traditional overlap analysis tools focus on full-text resources, making it challenging to assess the content of index and abstract databases in the context of resources which contain full text. We decided to develop our own tool in house both due to a lack of available funds to subscribe to an existing tool such as Gold Rush (http://www.coalliance.org/grinfo/) which offers a content analysis module, including indexes, and because the process would improve our programming skills—a benefit which will endure into the future.

Data Collection

Initially, we identified the type of data we would like to generate and present to the user; thereby, identifying the necessary raw data required for processing:

- Journal Title
- ISSN
- EISSN
- Coverage type (selective or full content for a given title)
- Start and stop dates for Indexing and Full-Text coverage
- Embargo details

We began looking at vendor websites for title lists containing the data we required. Some vendors, EBSCO and ProQuest, for example, provided detailed information for many databases. Some society publishers, however, did not provide data in an easily downloadable format or provided very limited details beyond titles and ISSNs. In some cases it was necessary to copy and paste data into a spreadsheet and then reformat it, either manually or programmatically, in order for it to be useful.

We then had to make decisions about the type of content to include in our analysis. We decided at the outset to only address periodical coverage, eliminating both e-books and "periodical/serial books" (with ISBNs rather than ISSNs). Additionally, when selective coverage of a given title was acknowledged in a database, we elected to upload it for future availability but not include it in the current reports. Selective coverage is used particularly in specialized subject databases whereby only select articles, sometimes occasional whole issues, of a periodical are

http://dx.doi.org/10.5703/1288284315104

included. Terminology regarding this concept varies by vendor, and one must be careful to understand the level of coverage in order to make consistent decisions across vendors. Once the data was gathered, we prepared the spreadsheets for automated upload by removing unneeded columns, ordering columns consistently, and removing column headings.

Data Normalization

As is the case with serials, notation for serial titles varies from data source to data source. Since we are attempting to compare titles across various data sources, making titles uniform is necessary. We encountered these main deviances from standards:

- Use of diacritics and special characters (e.g., fur vs. für; an ellipsis character vs. three periods)
- Capitalization (Some used all caps, others used title capitalization)
- Inconsistent punctuation (e.g., trailing or leading periods)
- Additional information in the title demarcated with various punctuation (e.g., translated titles, ISSNs embedded in the title)
- Additional spaces

The easiest thing to do with diacritics and special characters is to simply normalize them to be consistent with the ASCII character set. That is, remove them entirely. We constructed a translation table to substitute one or more ASCII character for a single diacritic. We were also able to make simple substitutions to solve other inconsistencies, such as:

- Transform "& "to "and"
- Enforce spacing around a colon (" : " replaced by " : ")

We used a number of regular expressions to remove other inconsistencies:

- Remove anything in parentheses
- Remove brackets surrounding an item; delete everything following the set of brackets. For example, this would change *[Hoigaku no jissai to kenkyu] [Study and practice of medical jurisprudence]* into *Hoigaku no jissai to kenkyu*
- Remove a trailing period and any trailing spaces
- Remove any leading periods or spaces
- Truncate a title at a slash
- Remove leading and trailing articles "la" and "the"
- Remove all ellipses and trailing/leading non-word spaces
- Replace multiple spaces with a single space

This normalization so far only deals with titles that are similar to each other and differ only in minor ways. We were aware of cases where a title had changed somewhat dramatically, yet it retained its former ISSN. Because of this, and because ISSNs are subject to fewer rules, we set about to normalize and ingest ISSNs as well. Normalizing ISSNs was a simple process:

- Ensure that there is a hyphen in the fifth position from the right
- Left-pad with zeros if there are fewer than nine characters

The impact of this normalization was substantial. While we ingested 100,908 unique (raw) titles, our normalization reduced that number to 55,792 normalized titles. Nearly 32,000 of the normalized titles only referred to a single raw title. Some 11,000 titles had two raw titles normalized together, and over 7,000 had three raw titles associated with them. Very few had more than ten raw titles normalized together.

Figure 1 reveals the spread of titles normalized together. When we examine the higher end of the data, we see that we encountered several problems. For example, at data point 62 below, our raw titles were associated with a single normalized title because they shared an ISSN. In this case, RILM gave us an ISSN of "0000-0000" for several (62, as it happens) different titles. Unfortunately, this caused our loader to treat them as though they were title changes. It also

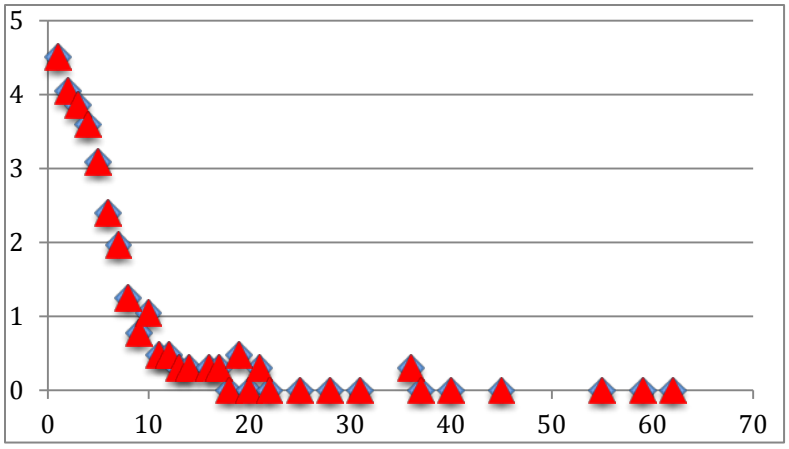


Figure 1. Scatter Graph of Raw Titles Normalized Together on a Logarithmic Scale

reduced the titles for that database by 30%. This is something that will be refined in future versions of the loader.

We considered normalizing the coverage data. Frequently, there was no coverage data, or vendors used different date notations. It also became apparent that we were merely displaying the date information, and we could rely on humans to parse the data. Therefore, we left the data as an unparsed string.

Database Architecture

We designed an SQL schema to capture the data and relationships that were important. At its core, we were only interested in a few items: databases, titles, and ISSNs. However, we needed a number of relationships to connect the three together. We also needed to connect raw titles to normalized titles. As Figure 2 indicates, these links introduce a significant amount of complexity.

Because our title lists were really an abstraction of the relationship of a database's relationship to a

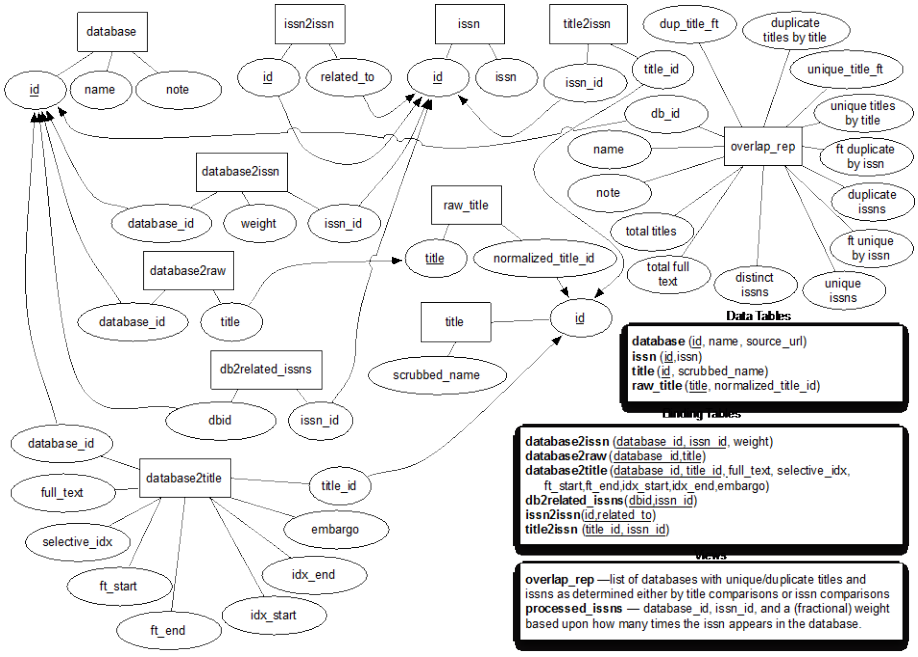


Figure 2. Database Entity-Relationship Diagram

certain title, we decided to attach our coverage information to the *database2title* binding table. This is where each title's coverage, indexing, and, embargo data is stored. This also presents a challenge when determining if a particular ISSN contains full text.

We also determined that there was a problem in counting ISSNs. In short, it is possible for a title to have multiple ISSNs. We found this several times. It was clear that several (usually two: a print ISSN and an electronic ISSN) ISSNs could refer to a single title. However, that should still be counted as a single title.

To solve this problem, we designed a weighting system to account for the multiple ISSNs. In short, we would determine how many ISSNs were related to each other and, therefore, a single title. The weight calculation itself is a simple formula. We would like all of the ISSNs that refer to a single title to end up equaling one, when we sum them together, so we simply divide 1 by the count of a title's ISSNs.

$$issn\ weight = \frac{1}{ISSN\ Count}$$

Figure 3. Formula to Determine an ISSN Weight

This necessitated creation of several tables: *issn2issn*, which relates ISSNs to each other, and *db2related_issns*. Each of them would be added to the *db2related_issns* table. Then we could create a weight for each ISSN. Initially we used a view, called *processed_issns* for this weight, but the query takes a substantial amount of time to run, so we incorporated it into the loading process and inserted the weight into *database2issn*.

Query Overview

When we embarked upon this project, we started with the results in mind. We wanted to clearly show collection managers the following fields for each database:

- Total titles
- Unique titles
- Duplicate titles
- Full text percentage for each data point

For our own curiosity, we wanted to know if it might be easier to rely upon ISSNs or normalized titles. Instead, we ended up using a hybrid approach. Neither ISSNs nor titles are regular enough to autonomously determine when a title is a title. However, in combination, we were successful at maintaining some links through title changes.

The lion's share of the work that went into the queries for this project went into the main overlap report, which can be seen in the top right corner of Figure 2. This query encompasses nearly 200 lines, 101 of which deals with rolling the ISSN data up as it relates to each database. The query takes approximately 17 seconds to run on our

Database Name	Collection Note	Total Titles	Unique by Issn	Duplicate by Issn	Unique by Title	Duplicate by Title
2012 Springer Journal Title List	Springer (EBSCO)	1741 FT 100.0%	44 (2.5%) FT 100.0%	1690 (97.5%) FT 100.0%	20 (1.1%) FT 100.0%	1721 (98.9%) FT 100.0%
ABIINform Complete	GALILEO	5531 FT 78.7%	1589 (28.6%) FT 84.9%	3958 (71.4%) FT 76.3%	1545 (27.9%) FT 85.2%	3986 (72.1%) FT 76.1%
Academic Search Complete	GALILEO	12686 FT 65.9%	1710 (13.4%) FT 88.5%	11048 (86.6%) FT 62.5%	1655 (13.0%) FT 88.3%	11031 (87.0%) FT 62.5%
Accounting and tax	GALILEO	570 FT 66.3%	110 (19.3%) FT 46.4%	460 (80.7%) FT 71.1%	107 (18.8%) FT 43.9%	463 (81.2%) FT 71.5%
ACM Digital Library	LYRASIS (Tarver)	136 FT 100.0%	58 (42.0%) FT 100.0%	80 (58.0%) FT 100.0%	57 (41.9%) FT 100.0%	79 (58.1%) FT 100.0%
Advanced Placement Source	GALILEO	5176 FT 100.0%	3 (0.1%) FT 100.0%	5212 (99.9%) FT 100.0%	2 (0%) FT 100.0%	5174 (100.0%) FT 100.0%
Alt Health Watch	GALILEO	139 FT 95.0%	27 (19.4%) FT 92.6%	112 (80.6%) FT 95.5%	25 (18.0%) FT 92.0%	114 (82.0%) FT 95.6%
America History and Life	Board of Regents (Tarver)	1691 FT 0%	275 (16.2%) FT 0%	1426 (83.8%) FT 0%	270 (16.0%) FT 0%	1421 (84.0%) FT 0%
American Chem Soc	Board of Regents (Tarver)	44 FT 100.0%	1 (2.3%) FT 100.0%	43 (97.7%) FT 100.0%	1 (2.3%) FT 100.0%	43 (97.7%) FT 100.0%

Figure 4. Master Overlap Report

commodity hardware, and the vast majority of that time is spent on the ISSN data as well. The title data is relatively straightforward, and the database2title relationship makes that very easy to get a hold of. Because the query takes so long, we insert the data into a table at load time to have a quickly retrievable home page. The first few rows of this report are shown in Figure 4.

Because we were interested in how well the ISSN-based data matched with the title-based data, we compared the two sets of unique and duplicate counts. Table 1 shows the standard deviation of the difference between the title-based and ISSN-based data. The mean difference was very low at 11.96 titles per database, which is less than a single percentage of the titles in the average database. The overall standard deviation between the various databases is fairly low, and the count was close to a single percentage, so we can conclude that although we did not have achieve perfection, we came acceptably close to our goal. There were two outliers in the unique fields: Article First and PubMed. In both cases, they had hundreds more unique ISSNs than the title count. Those caused the unique standard deviation to be a bit higher than the duplicate count.

	Count	Percentage
Unique Standard Deviation	44.52	1.28%
Duplicate Standard Deviation	13.23	1.08%
Total Standard Deviation	33.09	1.19%
Mean Difference	11.96	0.84%

Table 1. Comparing the Difference Between ISSN-Based and Title-Based Data

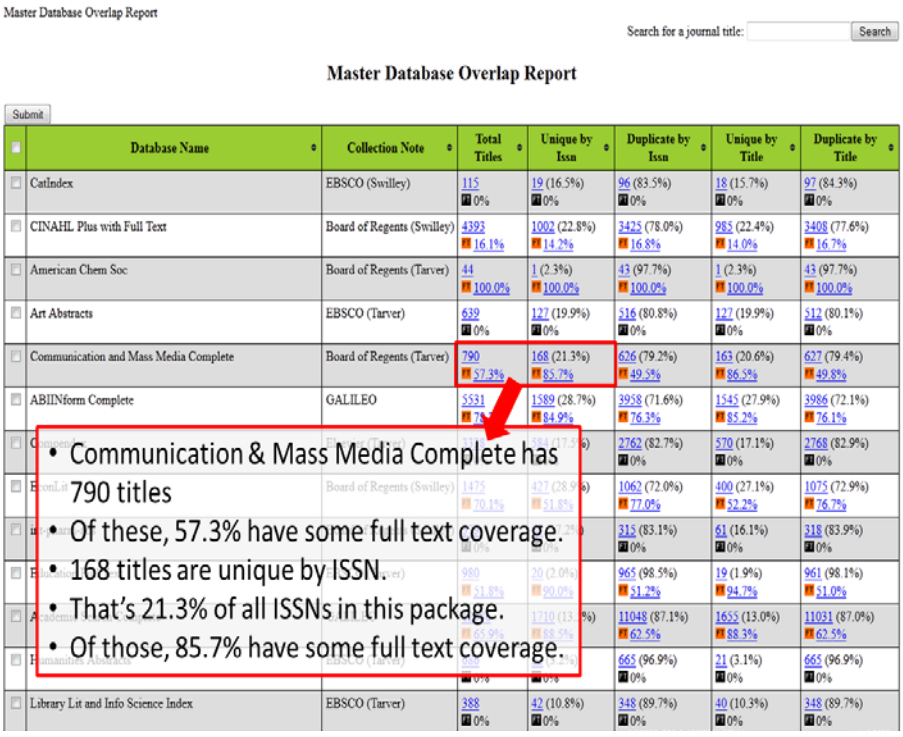
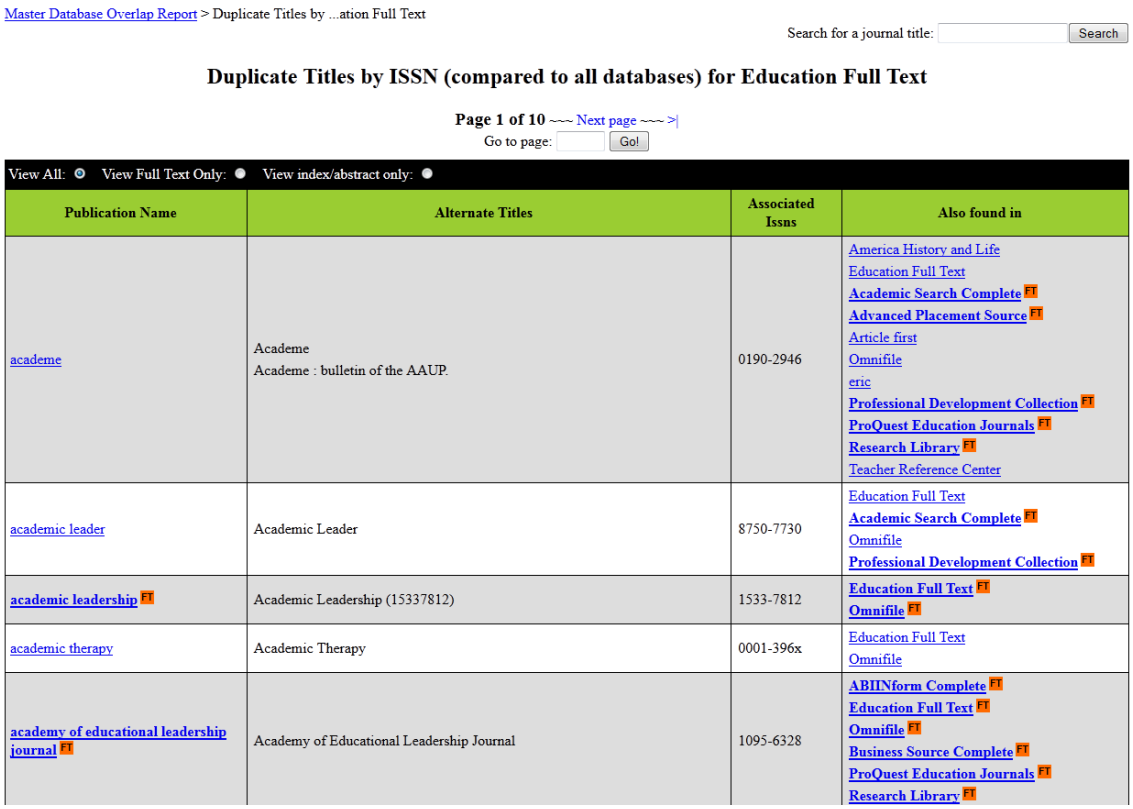


Figure 5. Master Database Overlap Report, Highlighting Communication and Mass Media Complete

The Duplicate Titles by ISSN report, shown in Figure 6, presents the list of titles where coverage is also found in other packages. The alternative databases are listed with an indication as to whether or not they contain any full-text coverage for that particular title. Alternate forms of the title, as found in the raw data, are listed for clarification and to help identify any problems where periodicals with similar titles might have been normalized during uploading to appear to become the same title.

The third tier of information is the title-specific report, as seen in Figure 7 for *Administrative Science Quarterly*. Here we see the list of packages which contain this title and, for clarification, we present the title as it was found in the raw data for each package, showing a full-text icon to identify any full-text content. This report shows any indexing or full-text coverage information provided by the vendor. An absence of data is acknowledged by the word 'None.' Where a cell for the end of indexing/full-text is empty, coverage is assumed to be ongoing. Users can also reach title-specific information at any time by conducting a title search in the search box in the upper right hand corner of the interface.

The interface is dynamically generated based upon each user decision. Master control is by JavaScript and jQuery which interacts with Python scripts on the CherryPy framework via AJAX and JSON. Query output is rendered as XML and then transformed into XHTML using XSL to present the report to the user. The interface features a breadcrumb trail to help the user return to previous queries. This trail is incrementally eliminated as one traverses backwards through the breadcrumbs. Additionally, due to the large volume of results for some queries, we also introduced pagination. The pagination functions by downloading all of a query's results into a session where it is speedily accessible. A set number of rows of data are presented at a time, and as long as the user remains within the same report context, the query is not rerun as the user navigates through the pages of results. This improves performance.



Master Database Overlap Report > Duplicate Titles by ...ation Full Text

Search for a journal title: Search

Duplicate Titles by ISSN (compared to all databases) for Education Full Text

Page 1 of 10 ~~~ Next page ~~~ >|

Go to page: Go!

View All: View Full Text Only: View index/abstract only:

Publication Name	Alternate Titles	Associated Issns	Also found in
academe	Academe Academe : bulletin of the AAUP.	0190-2946	America History and Life Education Full Text Academic Search Complete FT Advanced Placement Source FT Article first Omnifile eric Professional Development Collection FT ProQuest Education Journals FT Research Library FT Teacher Reference Center
academic leader	Academic Leader	8750-7730	Education Full Text Academic Search Complete FT Omnifile Professional Development Collection FT
academic leadership FT	Academic Leadership (15337812)	1533-7812	Education Full Text FT Omnifile FT
academic therapy	Academic Therapy	0001-396x	Education Full Text Omnifile
academy of educational leadership journal FT	Academy of Educational Leadership Journal	1095-6328	ABIINform Complete FT Education Full Text FT Omnifile FT Business Source Complete FT ProQuest Education Journals FT Research Library FT

Figure 6. Report Showing Duplicate Titles by ISSN For Education Full Text

Master Database Overlap Report > Duplicate Titles by ...ation Full Text > administrative science quarterly

Search for a journal title: Search

administrative science quarterly

View All: ◉ View Full Text Only: ○ View index/abstract only: ○

Database Name	Title as found in Database	Full Text Start	Full Text End	Full Text Embargo (Months)	Index Start	Index End
ABIINform Complete FT	Administrative Science Quarterly	03/01/1987	09/01/2001		06/01/1971	Current
Advanced Placement Source FT	Administrative Science Quarterly	01/01/1985			01/01/1985	
America History and Life	Administrative Science Quarterly	None	None		03/01/1964	None
Article first	Administrative science quarterly.	None	None		1980	2010
Business Source Complete FT	Administrative Science Quarterly	06/01/1956			06/01/1956	
Education Full Text FT	Administrative Science Quarterly	03/01/1995	03/01/2011		09/01/1981	None
Omnifile FT	Administrative Science Quarterly	03/01/1995	03/01/2011		09/01/1981	None
ProQuest Social Science Journals FT	Administrative Science Quarterly	03/01/1987	09/01/2001		06/01/1971	Current
PsycInfo	Administrative Science Quarterly	None	None		1956	None
PubMed	Administrative science quarterly	None	None		None	None
Research Library FT	Administrative Science Quarterly	03/01/1987	09/01/2001		06/01/1971	Current
Social Science Abstracts	Administrative Science Quarterly	None	None		09/01/1981	12/01/2002
Social Science Citation Index	ADMINISTRATIVE SCIENCE QUARTERLY	None	None		None	None
SocIndex with Full Text FT	Administrative Science Quarterly	06/01/1956	None		06/01/1956	None

Figure 7. Title Report Showing Coverage for *Administrative Science Quarterly*

Use

Subject liaisons and collection managers are the primary users of this tool. Initially, they identify high duplication packages and, in the context of their subject knowledge, they investigate the title lists to assess alternative coverage for a given title. Unique coverage is also assessed to determine if any key titles are included there and nowhere else. As this tool is introduced to public services librarians, we emphasize the importance of interpreting the data output of the system in context. The subject liaison must be aware both of key titles in the field and the nature of the e-content packages relevant to the field in order to benefit fully from the available reports.

The level of use of an electronic resource is a function not merely of uniqueness. The usability of the interface and the concentration of subject-relevant material as well as the education of library patrons by librarians also have an impact. It could be said that a unique but unused package is not as useful as a semi-unique but well-used package. Returning to the notion of assessing indexes, an index which contains information about key titles which is not available in any other package represents the only well-organized means by which that content can be discovered. The ongoing relevance of Interlibrary Loan operations hinges upon this.

Conclusion

It is indeed a worthwhile endeavor to reorganize, enhance, and supplement existing information to assist with decision making. The tool we have created is not perfect, but with it we have attained a sufficient standard to assist us with the process of making tough decisions, especially where there is a danger of targeting index/abstract-only resources as the default go-to place to make cuts. Perhaps the most lamentable aspect of the tool is the ever-changing content of title lists. If we don't keep it up to date with content, then it merely becomes a static waypoint in the history of e-resource availability.