ARTICLE

A Framework for Measuring Relevancy in Discovery Environments

Blake L. Galbreath, Alex Merrill, and Corey M. Johnson

ABSTRACT

Discovery environments are ubiquitous in academic libraries but studying their effectiveness and use in an academic environment has mostly centered around user satisfaction, experience, and task analysis. This study aims to create a quantitative, reproducible framework to test the relevancy of results and the overall success of Washington State University's discovery environment (Primo by Ex Libris). Within this framework, the authors use bibliographic citations from student research papers submitted as part of a required university class as the proxy for relevancy. In the context of this study, the researchers created a testing model that includes: (1) a process to produce machine-generated keywords from a corpus of research papers to compare against a set of human-created keywords, (2) a machine process to query a discovery environment to produce search result lists to compare against citation lists, and (3) four metrics to measure the comparative success of different search strategies and the relevancy of the results. This framework is used to move beyond a sentiment or task-based analysis to measure if materials cited in student papers appear in the results list of a production discovery environment. While this initial test of the framework produced fewer matches between researcher-generated search results and student bibliography sources than expected, the authors note that faceted searches represent a greater success rate when compared to open-ended searches. Future work will include comparative (A/B) testing of commonly deployed discovery layer configurations and limiters to measure the impact of local decisions on discovery layer efficacy as well as noting where in the results list a citation match occurs.

INTRODUCTION

Discovery environments are ubiquitous in academic libraries as all but two libraries in the Association of Research Libraries (ARL) report using a discovery environment, and they continue to gain traction in other library settings.[1] The one-stop shopping model of discovery environments is one of their most alluring features as it closely resembles searching the open web. This familiarity allows users who are accustomed to searching the web to feel comfortable searching the library catalog without fear of encountering a "failed" search (zero result set). Discovery environments seldom fail to return results as even the most rudimentary or naïve search strategy will return something for a user. This idea of "returning something" has been anecdotally noted as a positive as it ensures the user does not give up and allows novices to be successful with limited search sophistication or prior instruction from information professionals.

One of the potential negatives to this approach however is the sheer volume of material that is returned per search query. Library discovery environments often present thousands, if not millions, of search results from an initial search query. This emulation of Google is essentially

Blake L. Galbreath (blake.galbreath@wsu.edu) is Core Services Librarian, Washington State University. **Alex Merrill** (merrilla@wsu.edu) is Head of Library Systems and Technical Operations, Washington State University. **Corey M. Johnson** (coreyj@wsu.edu) is Instruction & Assessment Librarian, Washington State University.

https://doi.org/10.6017/ital.v40i2.12835

making the time-honored study of relevancy (precision/recall) moot. How can one determine the number of relevant documents in a search query if the number of documents returned is becoming limitless? This study aims to create a quantitative, reproducible framework to test the relevancy of results returned from, and the overall efficacy of, a library discovery environment, in this case, Ex Libris Primo. Within this framework, the authors compare the results returned in model Primo search queries against the bibliographic citations used in students' research papers.

BACKGROUND

The University Common Requirements (UCORE) curriculum, implemented in fall 2012, was a major redesign of the Washington State University (WSU) undergraduate general education program. UCORE is comprised of required categories of classes designed to build student proficiency in the Seven Undergraduate Learning Goals.[2] Roots of Contemporary Issues (RCI) is the sole mandated undergraduate course under the UCORE system.[3] During the 2018–2019 academic year, over 4,500 students were enrolled in RCI at WSU, the vast majority being first-year students.

This paper utilizes data from the RCI Library Research Project, a term-length research experience with four central assignments designed to familiarize students with the fundamentals of quality research and a cumulative research paper where they utilize the skills learned. The research project components are spaced evenly throughout the term; students are guided along the research process from general topic formation, to research question generation, to thesis statement defense in the final paper. Students are tasked with finding sources of particular resource types (e.g., journal articles), describing the value of these sources for their research, and citing them properly in Chicago Style.

WSU Libraries uses the discovery environment Primo, an Ex Libris product, to provide resources to its patrons.[4] Specifically, WSU Libraries uses the New User Interface version of Primo, which incorporates search results from the Primo Central Index (PCI) in its default search. Primo, like all discovery environments, provides results with a wide variety of resource types so RCI students can use it at all stages of the term research project. Students use it in the pursuit of contemporary newspaper articles, history monographs, history journal articles, and primary sources. In this article, the authors focus on the versatility of Primo, using RCI student paper bibliographies as the central data source for the project.

LITERATURE REVIEW

The need for assessment of library resources and services in higher education has been well-documented. Libraries are increasingly asked to provide tangible evidence they aid student information literacy skill development and thus advance achievement of institutional learning outcomes. Accrediting bodies acknowledge, "the importance of information literacy skills, and most accreditation standards have strengthened their emphasis on the teaching roles of libraries."[5] Oakleaf and Kaske also stress the importance of librarians choosing assessments that can contribute to university-wide assessment efforts, noting they are preferable to assessments that only benefit libraries.[6] The Washington State University Libraries is committed to assessment of its resources and services, with Primo as a central target resource, and with large, lower-undergraduate courses as a primary area of focus.

There are numerous papers which document usability testing of Primo. Prommann and Zhang (2015) analyzed the efficiency of Primo through hierarchical task analysis (HTA). They counted the number of physical and cognitive steps necessary to get to records or full text of known items and concluded that Primo is "a flexible discovery layer as it helps achieve many goals with minimum amount [*sic*] of steps."[7] Although many of these studies articulate avenues of success in terms of user interaction with the discovery environment, there are also reports of difficulties in a variety of categories. Students have problems with source retrieval, for example, understanding availability status terminology and labels, and using link resolvers and interlibrary loan.[8] Dalal et al. (2015) demonstrated that retrieving the full text of an article in a discovery environment is sometimes unintuitive for students and involves navigating multiple interfaces.[9] Users also have issues using facets to find particular resource types or distinguishing between them.[10] While the study addressed in this paper does not directly address user difficulties with Primo functionality, issues with source retrieval point to a plausible explanation for the few matches between the model search results and student paper bibliographies. It is possible students saw many of the same sources from the model searches in their results, but ultimately did not secure those sources because of the difficulties outlined above. In other words, some source selection choices are based mostly on availability, not as much on relevance.

Source relevancy is an active area of research for web-based discovery services, in terms of comparative studies to disciplinary subject databases. Evelhoch and Zebulin analyzed two years of usage data from both Primo and a selection of subject databases, concluding that users have difficulty finding relevant sources in Primo or they are not available.[11] Based on users' judgments, Lee and Chung, determined that EBSCO Discovery Service was less effective than a set of education and library subject databases in terms of source relevance.[12] Another study illustrated that while students preferred discovery environments, the articles they selected from the subject (indexing and abstracting) databases were more authoritative.[13] Finally, librarians are posited to believe that subject databases are superior to discovery environments in terms of the relevancy of search results and disciplinary coverage.[14] Conclusions about source relevancy are complicated by the fact that students infrequently look beyond a first page of results lists.[15]

Researchers have also explored the idea of Primo user satisfaction through the presence of relevant results. In one instance, using online questionnaires and in-person focus groups, researchers found users had a high level of satisfaction with their institution's discovery environment, largely attributed to the quality of search results over ease of use.[16] Hamlett and Georgas (2019) conducted a mixed-methods user experience study to understand student perceptions of relevancy in Primo. This study found that participants believed Primo to return relevant results (with an average score of 8.3 out of 10). However, some of the qualitative responses indicated that the keywords used did not actually yield relevant results.[17] Many other methods and measures have been executed in determining the value and usefulness of Primo. Huurdeman, Aamodt, and Heggo analyzed a dataset of 50 popular queries in Primo. They deemed a query successful if the first 10 results included the (likely) targeted resource and found that 58% of the queries from the popular searches dataset had been successful, while 20% were unsuccessful, and 22% could not be determined. Their approach assumed there is one intended document per query and that the authors can surmise what it is.[18] The research presented in the remainder of this article below is unique in that the authors explore user judgment of source relevance (satisfaction) as a function of whether sources in the model Primo searches for their topics existed in the students' papers' bibliographies.

METHODS

Research Questions

The impetus for this study was to understand the factors that play a role in establishing a framework to test the relevancy of results returned from Primo. The authors attempted to answer the following questions:

- How effective is Primo at returning relevant results?
- To what extent does faceting improve search results?
- Which search strategies are the most effective within the given framework?
- How can the researchers refine the framework for future investigations into relevancy?
- What are the implications of this study for end users?

Data Collection

The authors began with a sample of 100 randomly selected and anonymized research papers that were submitted to the Roots of Contemporary Issues (RCI) courses in fall 2018 and spring 2019 semesters. The study used a two-pronged approach to generate keywords for model Primo search queries. For one approach, keywords were machine-generated via a word-vector generation process. For the other, keywords were human-generated by a student research assistant to approximate natural language queries.

Keywords and Queries, Machine

A RapidMiner (https://rapidminer.com/) word-vector generation process with term-frequency schema converted the research papers into keywords, which the authors then used to generate search queries. Within the main routine, the Process Documents from Files operator, RapidMiner transformed the texts into lower case and tokenized the final papers according to non-letters. RapidMiner then filtered the data by those tokens representing nouns and adjectives, removed English stop words, and filtered tokens by length, with a minimum of one character and maximum of 50 characters. The researchers then applied a Snowball stemmer for English words and generated 20 n-grams per paper, each with a maximum length of four. Table 1 illustrates the product of the word-vector generation process. Throughout this example research paper, "trade" occurred 40 times, "slave" occurred 34 times, "slave" and "trade" occurred together 26 times, "africa" occurred 18 times, "impact" occurred 16 times, "african" occurred 11 times, and "peopl" occurred 10 times.

Table 1. Example N-grams and frequency as retrieved from RapidMiner

N-gram	Number of occurrences
trade	40
slave	34
slave_trade	26
africa	18
impact	16
african	11
peopl	10
...	...

Number of N-grams

After compiling the data in RapidMiner, the authors created a process to select those n-grams to use in the model Primo search queries. Huurdeman, Aamodt, and Heggo (2018) found that users included an average of 2.6 terms per query in their Popular Searches dataset.[19] In a report by Ex Libris, Stohn indicates that most topic-search queries contain five or fewer words.[20] In order to investigate both ends of this spectrum, this study constructed short-length queries, consisting of two n-grams, and full-length queries, consisting of four n-grams, using the following rubric to help systematize the construction.

Rubric to Select N-grams for Short- and Full-Length Queries

Pick terms that satisfy the following criteria:

1. N-grams that occur more frequently in a paper are preferred to those that occur less frequently.
2. If two n-grams appear to be structural derivatives of the same word (e.g., korea and korean), select the shortest n-gram and truncate it.
3. If one or more of the top terms appear in a later 2-gram, use the 2-gram as a phrase search.
4. Ignore n-grams with repeating terms (e.g., south_africa_africa).
5. Truncate all terms (using asterisk or question mark), except the first term of a phrase search, unless the first term is not a complete word (e.g., "busi* meeting*").
6. For terms or phrases that end in truncated "i", use the truncated version of the term and its truncated "y" counterpart, and combine both with an OR operator (e.g., countri* OR country*).
7. Ignore all 3- and 4-grams as they have a propensity to create nonsensical phrase searches (e.g., racism_polic_brutal).
8. If abbreviations are encountered, expand them for searching purposes (e.g., US is "united states"), except in cases where they are more commonly known by their abbreviation (e.g., ddt).
9. Ignore results of contractions (e.g., 't)

In case of a tie in the selection of an n-gram, sequence the following rules for selection:

1. Preference proper nouns over other nouns and adjectives. If there are multiple proper nouns, preference place-name proper nouns over other proper nouns.
2. Preference the n-gram that occurs in the greatest number of two or more n-grams later in the list.
3. Preference longer words over shorter words.
4. Group all the tied n-grams with a series of OR statements. Note: this may result in the selection of more than four total n-grams.

Referring to the example n-grams from table 1, an illustration of this method is shown in the following steps:

1. Arrange terms from highest to lowest frequency.
2. Select slave_trade as first n-gram, since "trade" and "slave" both occur in later n-gram. Truncate to "*slave trade**".
3. Select *africa* since it has the next greatest number of occurrences. Combine *africa* with *african* since they are structural derivatives of one another. Truncate to *africa**.*

At this point, the first two selected n-grams—slave_trade and africa—become the keywords of the short-length query *"slave trade*" AND africa**.

4. Select *impact* since it has the next greatest number of occurrences. Truncate to *impact*.*
5. Select *peopl* since it has the next greatest number of occurrences. Truncate to *peopl*.*

Finally, the first four selected n-grams—slave_trade, africa, impact, and peopl—become the keywords of the full-length query *"slave trade*" AND africa* AND impact* AND peopl**. On average, after stop words and Booleans were removed, the full-length queries in this study were 5.69 keywords long, while the short-length queries were 3.11 keywords long.

Keywords and Queries, Natural Language

In addition to the machine-oriented keyword process, the authors employed a student research assistant to create human-generated phrases, consisting of 3–10 words, which served as synopses for each of the 100 papers. This study then used these phrases as proxies for creating natural language search queries. For the same example research paper cited in table 1 above, this student created the summary phrase *history and effects of the slave trade*. This phrase in its entirety became the natural language query. On average, after stop words and Booleans were removed, the natural language queries used in this study were 3.95 keywords long.

Search Results

Using the three keyword-generation strategies outlined above, the authors constructed search queries and ran them against the Ex Libris' Primo Search API endpoint. Table 2 summarizes example result sets from the above short-length query, full-length query, and natural language query.

For each of the keyword-generation strategies, the authors constructed search queries along four parameters: queries that used no faceting (open-ended), queries that faceted to articles only (articles), queries that faceted to books and ebooks only (books), and queries that faceted to newspaper articles only (newspapers). In all, there were 12 search-query constructions (three query types by four faceting modes) for fall 2018 and 12 for spring 2019. To construct a baseline for the search comparisons, the researchers designed the initial search to be open-ended. That is, the study assumed that patrons most often use the default, basic search functionality, with no facets selected. A segment of the RCI instruction specifically encourages students to incorporate materials with resource types articles, books, and newspaper articles into their research papers. The authors therefore assumed that these students would most likely utilize facets corresponding to these resource types in their more specific queries and mirrored this behavior in the comparative searches. Each Primo Search API returned titles for the top 50 results, moving beyond users' usual search behavior in an effort to provide more flexibility to the initial steps of the relevancy framework.

Table 2. First-occurring result titles for query types: Short-length, full-length, and natural language queries

Query type	Query	First-occurring result titles
Short-length	"slave trade*" AND africa*	The Atlantic slave trade The Atlantic slave trade : a census The Atlantic slave trade Legacy of the trans-Atlantic slave trade : hearing before the Subcommittee on the Constitution, Civil Rights, and Civil Liberties of the Committee on the Judiciary, House of Representatives, One Hundred Tenth Congress, first session, December 18, 2007. ...
Full-length	"slave trade*" AND africa* AND impact* AND peopl*	The Atlantic slave trade The Atlantic slave trade : effects on economies, societies, and peoples in Africa, the Americas, and Europe Slave trades, 1500–1800 : globalization of forced labour African voices of the Atlantic slave trade : beyond the silence and the shame ...
Natural-language	history and effects of the slave trade	Urban History, the Slave Trade, and the Atlantic World 1500–1900 The Atlantic slave trade and British abolition, 1760–1810 The Decolonization of African Education and History The United States and the transatlantic slave trade to the Americas, 1776–1867 ...

A student research assistant harvested all the citations used across the 100 example papers to create an inventory of 730 bibliographic citations. Using the Excel Fuzzy Lookup Add-In, the authors then compared this bibliographic inventory against the 60,000 titles that were returned

via the Primo Search API. This add-in fuzzy matches rows between two different tables and assigns a similarity score for each match. The study focused attention on rows with matching scores of .80 and above to further investigate potential matches. Using the fuzzy matches as a starting point, the authors confirmed or denied matches by hand, using title and resource type as the main criteria.

Table 3. Sample comparison of citations used in research papers against results returned from Primo search API

Fuzzy score	Citation title	Citation resource type	Results title	Result resource type	Confirmed match
1.0000	A Short History of Biological Warfare	Article	A short history of biological warfare	Article	Yes
0.9933	THE FEMALE MADLADY Women, Madness, and English Culture, 1830–1980	Print book	The female malady : women, madness, and English culture, 1830–1980	Print book	Yes
0.9778	Industrial Revolution	Web resource	The industrial revolution	e-book	No
0.9037	Drug Use & Abuse	Print book	Drug use and abuse : a comprehensive introduction	Print book	No

RESULTS

Source Citation Data Description

This study compared citations gathered from a random sample of 100 research papers from the two semesters of all sections of History 105/305 taught at Washington State University (WSU) from fall 2018 to spring 2019. Table 4 below gives a descriptive breakdown of the citations by resource type. The student research assistant identified and categorized the source citation list.

Table 4. Total source citations

Resource type	Fall 2018 (% of total)	Spring 2019 (% of total)
Book chapter	7 (1.94%)	4 (1.08%)
Books (e-books/print)	107 (29.72%)	96 (25.95%)
Newspaper article	63 (17.50%)	60 (16.22%)
Journal article	84 (23.33%)	99 (26.76%)
Reference entry	6 (1.67%)	6 (1.62%)
Other/ Cannot determine	10 (2.78%)	15 (4.05%)
Web document	81 (22.50%)	90 (24.32%)
Magazine article	1 (.28%)	N/A
Newspaper/Magazine article	1 (.28%)	N/A
Semester citation count	360 (100%)	370 (100%)
Total citation count	730	

Target Citations List Data

The citations collected from the papers were then compared against 60,000 citations retrieved from the WSU Primo Search API endpoint on July 24, 2020, as described previously in the methods section.

To better account for the differing numbers of citations among resource types in the source data and to normalize reporting across query types and semesters, most results are presented as a percentage and referred to as the *matching success rate*. For example, the natural language query had six matches out of a possible 360 citations in the open-ended search for citations from the fall of 2018. The matching success rate of the open-ended search in the fall of 2018 therefore is calculated at 1.67% (see table 5). Table 6 below shows the percentage results for short queries, and table 7 for full queries. For information about the raw source numbers and target data, please see the Open Science Framework project site.[21]

When all query types and faceting modes are considered, the matching success rate almost uniformly increased from fall 2018 to spring 2019. The largest difference in matching success rate was observed in the full-query articles only search at 8.91% as shown in table 7. The open-ended search observed the smallest difference in positive movement and the anomaly of a diminishing success rate. Across the natural language and full-query types the open-ended search exhibited the least amount of positive difference in success rate, at 1.04% and 0.26% respectively, and the short-query open-ended search had a small negative change in success rate at −0.36%.

Table 5. Natural language query results success rate

	Fall 2018	Spring 2019	% Difference
Open-ended search	1.67%	2.70%	1.04%
Articles only	4.76%	9.09%	4.33%
Books only	3.74%	11.46%	7.72%
Newspapers only	0.00%	1.67%	1.67%

Table 6. Short-query results success rate

	Fall 2018	Spring 2019	% Difference
Open-ended search	3.33%	2.97%	−0.36%
Articles only	3.57%	5.05%	1.48%
Books only	9.35%	10.42%	1.07%
Newspapers only	0.00%	3.33%	3.33%

Table 7. Full-query results success rate

	Fall 2018	Spring 2019	% Difference
Open-ended search	0.56%	0.81%	0.26%
Articles only	1.19%	10.10%	8.91%
Books only	0.93%	5.21%	4.27%
Newspapers only	0.00%	5.00%	5.00%

Total Unique Matches

Across all three search strategies and their four iterations, the researchers also note a raw count of matches which helps to determine how an overall search strategy is performing at finding matching citations. As the reader might expect, this metric includes a matching citation once across all four iterations of a search strategy. Meaning, even if a source citation appears in both the open-ended search and the books only search, that source citation is only counted once for the purpose of this metric.

For example, in the natural language query in fall 2018, six citations were matched in the open-ended search. Four of the citations were articles and two were books. Some of the matches in the articles and books searches were redundant to the open-ended search. Considering only unique matches in the articles, books, and newspaper searches, the authors calculated the total number of unique matches. When the target searches were compared, the researchers matched two additional citations in the books only citations list. When the authors add the two additional matches, there were a total of eight unique citation matches across all iterations of the natural language search (open-ended search, books only, articles only, newspapers only). The total unique matches number and the corresponding success rate of the total unique matches for each search strategy is shown in Table 8.

Table 8. Total unique matches

	Fall 2018	Spring 2019	% Difference
Natural language query	8 (2.22%)	22 (5.95%)	3.72%
Short query	14 (3.89%)	16 (4.32%)	0.44%
Full query	3 (0.83%)	18 (4.86%)	4.03%

Matches Added by Faceting

Another metric used to measure overall effectiveness of faceted searching is the percentage of matching citations that are new to the results list when limited to a certain resource type—*matches added by faceting*. Meaning, what matching citations were not present in the open-ended search results but are then matched when the results list is reduced to only a single resource type. In table 9, the percentage of matches that are new and only to be found in a targeted search result varies greatly. Between both semesters and among all search iterations, the smallest percentage of matches added by faceting is 14.29% and the largest is 83.33%.

Table 9. Matches added by faceting

	Fall 2018	Spring 2019	% Difference
Natural language query	2 (25.00%)	12 (54.55%)	29.55%
Short query	2 (14.29%)	5 (31.25%)	16.96%
Full query	1 (33.33%)	15 (83.33%)	50.00%

Comparing Search Strategies

The matching success rate across search strategies (natural, short, full) and iterations is a mixed result and does not allow for very useful comparison beyond descriptions of difference which are outlined in the comparison tables (tables 5–7). To better compare the search strategies as a whole, as opposed to how a particular iterative search performed relative to another open or targeted search, the researchers used a *weighted success rate* of the total unique matches from both semesters as the proxy for overall performance and the point of comparison among the three search strategies. The comparison of this weighted success rate shows no difference in overall success rate between the natural language query (4.11%) and the short query (4.11%). The search strategy that was demonstrably different in weighted success rate is the full query at a lagging 2.88%. See table 10 for comparison and calculation details.

Table 10. Weighted success rate of total unique matches

Natural language query	(2.22%*360)+(5.95%*370)/730	4.11% (0.04109589)
Short query	(3.89%*360)+(4.32%*370)/730	4.11% (0.04109589)
Full query	(0.83%*360)+(4.86%*370)/730	2.88% (0.02876712)

DISCUSSION

How Effective is Primo at Returning Relevant Results?

According to the preliminary findings, Primo is relatively ineffective at providing search results that match the citations used by the student researchers. The matching success rates of the open-

ended searches range from 0.56% to 3.33%. The possible reasons for these low numbers are numerous and varied; everything from students perhaps intending to use sources in the researchers' auto-generated results lists, but unfortunately were unable to locate the full text, to the prevalence of finding open internet sources outside the discovery layer, to open-ended searches being flooded with rarely cited reference materials and very contemporary newspaper articles (see more about these ideas below). Future research aims to understand more clearly which potential factors are present and to what degree they impact the matching success rates.

To What Extent Does Faceting Improve Search Results?

Faceting within Primo leads to better results, although the matching success rates are still more ineffective than not. The faceted searches contain the only matching success rates above ten percent: 10.10% (full query, articles only), 10.42% (short query, books only), and 11.46% (natural language query, books only). The data shows that the majority of unique matches found by the 2019 full-length and natural language search strategies occurs within the faceted searches (83.33% and 54.55%, respectively). It is interesting to note that these represent the two longer query strings, on average. Future testing will reveal whether there is a relationship between query length and percentage of matches added by faceting.

Which Search Strategies Are the Most Effective within the Given Framework?

Looking at the search strategies holistically, the researchers note that the total unique matches increased from fall 2018 to spring 2019 across all three query types. This increase was expected behavior, partially due to the fact that Primo relevancy ranking algorithms assume that patrons prefer newer materials.[22] The weighted success rate is an attempt to understand each search strategy's performance over the 2018–2019 academic year, as opposed to comparing one semester to the other. From this metric, the consistency of the short-length query is equally effective as the more dynamic performance of the natural language query. The researchers are looking forward to adding more data to this metric to understand in which direction the average might move.

How To Refine the Framework for Future Investigations into Relevancy

The most popular resource types used in the source citations were books, journal articles, web documents, and newspaper articles. Together, these categories comprised approximately 93% of all resource types in both fall 2018 and spring 2019. However, not all areas were equally accessible within Washington State University's discovery layer configuration. The heavy reliance on web documents in the source citations was somewhat problematic, given the fact that web documents did not constitute a faceted resource type in WSU Libraries' Primo prior to this study. Therefore, the authors will need to better account for web documents in future testing.

The assessment of newspaper articles also proved to be problematic, given their proclivity to inundate Primo search results with numerous and recent documents. The sheer number of newspaper articles published and indexed every year in Primo for general and introductory topics can dilute the pool of possible target citations greatly. For example, a scan of the matching newspaper articles reveals that 67% (4/6) were published in 2018. In future studies, the researchers will limit publication dates for target citations to the appropriate time period (e.g., an upper limit of May 2019 would be placed on publication dates for papers written in spring 2019) or collect data closer to the submission of research papers. In 11 out of 12 cases, matching success rates were better in spring 2019 than fall 2018, most likely due to recency. It is common for discovery environments, and true for the environment used in this study, to present content

sorted by relevance and then publication date. Therefore, the researchers expected to and did find an increased matching success rate closer to the date of testing, with the one exception of the short-length, open-ended search query.

This anomaly led researchers to dig more deeply into the target citations to see if a cause could be determined. Researchers found a larger than expected number of citations for resource types that are underrepresented in source citations. For example, the reference entry resource type surfaced prominently in the open-ended search for several of the queries, diluting the pool of target citations with entries that had little chance of appearing in the source citation lists. In one standout example, there were four separate reference entries titled simply "Taiping Rebellion." The discovery environment gave preference to these four separate reference entries over other, more substantive works, that are more likely to be cited in an academic paper. The researchers surmise this is partly a function of the relevancy ranking algorithm that gives greater weight to matches in the title, author, and subject fields.[23] Depending on the search and the configuration of the discovery environment, it is possible that reference entries would push other results from books, articles, and newspaper resource types farther down the results list, making them less and less visible in an open-ended search for a given topic. This dilution of the target citations with resource types that are not emphasized or widely used in source citations is another area the researchers aim to isolate and examine in further rounds of testing.

In addition to the source recency and particular source type issues explained above, the authors did not take into account source availability, nor where sources were found by students, which remains a confounding factor on matching success rate. Subsequent studies will capture whether sources are present in the local deployment of Primo during the time frame the students were conducting research. This issue will be further addressed and mitigated by analyzing URLs provided within student source citations.

Implications of This Study for End Users

The matching success rate in the open-ended search when compared to the type-limited searches leads to a discussion of how to define and present the default search of the discovery environment to best serve an academic population. More pointedly, it opens the discussion of what resource types to include within that default search to return the most relevant and useful results and not just the most results. In this case, the argument could be made that excluding several resource types (e.g., reference entries) would surface resources that are more likely to be cited in a researcher's scholarship.

Based on the number of matches that were introduced by performing a faceted search, it is evident that researchers still need to utilize a search strategy which includes using search filters and limiters (prior to or following the initial search) and other search tactics in a discovery environment to return relevant results. The notion that an open-ended "one and done search," for even the most introductory of topics, will be successful in retrieving many usable and citable resources in the first page or two of results is not supported by the results of this study.

CONCLUSIONS AND NEXT STEPS

As the common adage goes, "it's not what you say, it's what you do." In this study, the saying applies as the researchers move beyond what sources students think are relevant to the sources students ultimately use in their papers. The current slate of discovery environment research projects focuses largely on users' affective connections to discovery environments, often

compared to other kinds of academic databases, and places users in temporary, hypothetical research scenarios in order to judge source relevance.[24] In juxtaposition, the RCI research project is a term-length (10–14 weeks) venture; students have a significant amount of time and the aid of a scaffolded set of assignments, to bolster their source relevance assessment skills and authority. Methodologies which closely mirror the authentic experiences and curriculum of the students are those which arguably will provide a more accurate picture of the value of the discovery environment in an academic setting.

The authors of this study took the first steps in building a relevancy rating system for discovery environments. To standardize their preliminary results, they generated four metrics: matching success rate, total unique matches, matches added by faceted search, and weighted success rate. While the results of this study do not allow the researchers to draw statistical conclusions regarding the dominance of one search strategy over another in returning relevant results, the frequencies showed a better match (success) rate with faceted than non-faceted searching. Discovery environments are commonly advertised as providing an easy to use, one-stop location for academic research needs, but the reality is more complex. Students need to engage these systems with multiple search refinements to find valuable materials.

This investigation was also the initial attempt to create a machine-generated framework to test the relevancy of web-based discovery environment's results. As the authors look to build upon this preliminary study, there are several avenues to pursue that will enhance the methodology of the framework. One avenue is a refinement of the boundaries of the testing framework. This boundary refinement includes a re-examination of the criteria for inclusion in both the source citations and the search results list. In the current study, all student citations were deemed viable regardless of whether the source citation was able to be verified and accessed. This led to the inclusion of citations of lecture notes and other such materials that are not generally expected to appear in a discovery environment. The authors will also re-examine the inclusion of newspapers and reference works in open-ended searching. These two resource types are large in number, are not indexed very well, and often do not have descriptive titles. A portion of the next round of research will be dedicated to comparative testing (A/B) of generally deployed discovery environment configurations. Another avenue of exploration is determining where in the results list a citation appears, not just the binary positive or negative, and measuring any impact based on behavior of the search (i.e., search construction) or behavior and configuration of the discovery environment. Refining the methodology of the current framework will result in fewer potentially confounding factors and allow librarians to regain an understanding of relevancy when it comes to teaching discovery layers to student researchers. These next steps will contribute to the overall picture concerning the value and efficacy of web-based discovery environments that is steadily taking shape.

ENDNOTES

[1] Marshall Breeding, "Library Technology Guides: Academic Members of the Association of Research Libraries: Index-Based Discovery Services," Library Technology Guides, https://librarytechnology.org/libraries/arl/discovery.pl.

[2] "Student Learning Goals," Washington State University Common Requirements, 2018, https://ucore.wsu.edu/about/learning-goals.

[3] "Welcome to the Roots of Contemporary Issues," Washington State University Department of History, 2017, https://ucore.wsu.edu/faculty/curriculum/root/.

[4] "Search It," Washington State University Libraries, 2020, https://searchit.libraries.wsu.edu/.

[5] Megan Oakleaf and Neal Kaske, "Guiding Questions for Assessing Information Literacy in Higher Education," *portal: Libraries and the Academy* 9, no. 2 (2009): 277, https://doi.org/10.1353/pla.0.0046.

[6] Oakleaf and Kaske, "Guiding Questions."

[7] Marlen Prommann and Tao Zhang, "Applying Hierarchical Task Analysis Method to Discovery Layer Evaluation," *Information Technology and Libraries* 34, no. 1 (2015): 97, https://doi.org/10.6017/ital.v34i1.5600.

[8] Rice Majors, "Comparative User Experiences of Next-Generation Catalogue Interfaces," *Library Trends* 61, no. 1 (2012): 186–207, https://doi.org/10.1353/lib.2012.0029; David Comeaux, "Usability Testing of a Web-Scale Discovery System at an Academic Library," *College & Undergraduate Libraries* 19, no. 2–4 (2012): 199, https://doi.org/10.1080/10691316.2012.695671; Greta Kliewer et al., "Using Primo for Undergraduate Research: A Usability Study," *Library Hi Tech* 34, no. 4 (2016): 566–84, http://doi.org/10.1108/LHT-05-2016-0052; Blake Galbreath, Corey M. Johnson, and Erin Hvizdak, "Primo New User Interface," *Information Technology and Libraries* 37, no. 2 (2018): 10–33, https://doi.org/10.6017/ital.v37i2.10191.

[9] Heather Dalal, Amy Kimura, and Melissa Hofmann, "Searching in the Wild: Observing Information-Seeking Behavior in a Discovery Tool" (Association of College & Research Libraries 2015 Conference Proceedings, March 25–28, 2015): 668–75, http://www.ala.org/acrl/sites/ala.org.acrl/files/content/conferences/confsandpreconfs/2015/Dalal_Kimura_Hofmann.pdf.

[10] Comeaux, "Usability Testing"; Xi Niu, Tao Zhang, and Hsin-liang Chen, "Study of User Search Activities with Two Discovery Tools at an Academic Library," *International Journal of Human-Computer Interaction* 30, no. 5 (2014): 422–33, https://doi.org/10.1080/10447318.2013.873281; Kevin Patrick Seeber, "Teaching 'Format as a Process' in an Era of Web-Scale Discovery," *Reference Services Review* 43, no. 1 (2015): 19–30, https://doi.org/10.1108/RSR-07-2014-0023; Kylie Jarret, "Findit@Flinders: User Experiences of the Primo Discovery Search Solution," *Australian Academic & Research Libraries* 43, no. 4 (2012): 278–99, https://doi.org/10.1080/00048623.2012.10722288; Aaron Nichols et al., "Kicking the Tires: A Usability Study of the Primo Discovery Tool," *Journal of Web Librarianship* 8, no. 2 (2014): 172–95, https://doi.org/10.1080/19322909.2014.903133;

Kelsey Renee Brett, Ashley Lierman, and Cherie Turner, "Lessons Learned: A Primo Usability Study," *Information Technology and Libraries* 35, no. 1 (2016): 7–25, https://doi.org/10.6017/ital.v35i1.8965; Galbreath, Johnson, and Hvizdak, "Primo New User Interface."

[11] Zebulin Evelhoch, "Where Users Find the Answer: Discovery Layers Versus Database," *Journal of Electronic Resources Librarianship* 30, no. 4 (2018): 205–15, https://doi.org/10.1080/1941126X.2018.1521092.

[12] Boram Lee and EunKyung Chung, "An Analysis of Web-scale Discovery Services from the Perspective of User's Relevance Judgement," *Journal of Academic Librarianship* 42 (2016): 529–34, https://doi.org/10.1016/j.acalib.2016.06.016.

[13] Sarah P. C. Dahlen and Kathlene Hanson, "Preference vs. Authority: A Comparison of Student Searching in a Subject-Specific Indexing and Abstracting Database and a Customized Discovery Layer" *College & Research Libraries* 78, no. 7 (2017): 878–97, https://doi.org/10.5860/crl.78.7.878.

[14] Stefanie Buck and Christina Steffy, "Promising Practices in Instruction of Discovery Tools," *Communications in Information Literacy* 7, no. 1 (2013): 66–80, https://doi.org/10.15760/comminfolit.2013.7.1.135; Anita K. Foster, "Determining Librarian Research Preferences: A Comparison Survey of Web-Scale Discovery Systems and Subject Databases," *Journal of Academic Librarianship* 44 (2018): 330–36, https://doi.org/10.1016/j.acalib.2018.04.001.

[15] Diane Cmor and Xin Li, "Beyond Boolean, Towards Thinking: Discovery Systems and Information Literacy," 2012 IATUL Proceedings, paper 7, https://docs.lib.purdue.edu/iatul/2012/papers/7/; Kliewer et al., "Using Primo"; Alexandra Hamlett and Helen Georgas, "In the Wake of Discovery: Student Perceptions, Integration, and Instructional Design," *Journal of Web Librarianship* 13, no. 3 (2019): 230–45, https://doi.org/10.1080/19322909.2019.1598919.

[16] Courtney Lundrigan, Kevin Manuel, and May Yan, "'Pretty Rad': Explorations in User Satisfaction with a Discovery Layer at Ryerson University," *College & Research Libraries* 76, no. 1 (2015): 43–62, https://doi.org/10.5860/crl.76.1.43.

[17] Hamlett and Georgas, "In the Wake of Discovery."

[18] Hugo C. Huurdeman, Mikaela Aamodt, and Dan Michael Heggo, "'More Than Meets the Eye'—Analyzing the Success of User Queries in Oria," *Nordic Journal of Information Literacy in Higher Education* 10, no. 1 (2018): 18–36, https://doi.org/10.15845/noril.v10i1.270.

[19] Huurdeman, Aamodt, and Heggo, "More Than Meets the Eye."

[20] Christina Stohn, "How Do Users Search and Discover?: Findings from Ex Libris User Research," Ex Libris, 2015, https://www.exlibrisgroup.com/blog/ex-libris-user-studies-how-do-users-search-and-discover/.

[21] Alex Merrill and Blake L. Galbreath, "A Framework for Measuring Relevancy in Discovery Environments," 2020, https://osf.io/ve3kp/.

[22] "Primo Search Discovery: Search, Ranking, and Beyond," Ex Libris, 2015, https://www.exlibrisgroup.com/products/primo-discovery-service/relevance-ranking/.

[23] "Primo Search Discovery," 3.

[24] Lee and Chung, "An Analysis of Web-Scale Discovery Services"; Dahlen and Hanson, "Preference vs. Authority"; Lundrigan, Manuel, and Yan, "Pretty Rad"; Hamlett and Georgas, "In the Wake of Discovery."