Issues in Science and Technology Librarianship

A Comparison of Selected Bibliographic Database Search Retrieval for Agricultural Information

Stephanie M. Ritchie
Agriculture & Natural Resources Librarian
University of Maryland College Park
College Park, Maryland
sritchie@umd.edu

Kelly M. Banyas
Research & Instruction Librarian
The University of Scranton
Scranton, Pennsylvania
kelly.banyas@scranton.edu

Carol Sevin
Academic Services Librarian
Kansas State University
Manhattan, KS
sevin@ksu.edu

Abstract

Search result retrieval was compared across eight research literature databases (AGRICOLA, AGRIS, BIOSIS, CAB Direct, FSTA, Google Scholar, Scopus, and Web of Science) for three topics from different agricultural disciplines to compare retrieval results based on searcher experience. Precision, recall, and uniqueness were analyzed by rating search results (~2,400 citations) for relevancy. A generalized linear model statistical analysis determined that AGRICOLA ranked highest for precision and was statistically more likely to produce a relevant result than four other databases. CAB and Web of Science ranked highest for recall and both overlapped with AGRICOLA for statistical likelihood of producing a relevant result. Google Scholar retrieved the most unique content, but almost half of that content was not judged relevant. AGRICOLA, BIOSIS and CAB retrieved the most unique and relevant content. This study will help researchers and librarians working in the agricultural disciplines to select the bibliographic databases that will provide the most relevant search results and are most likely to meet their research need. It may also serve as a template for future bibliographic research in other disciplines.

Introduction

Bibliographic databases serve different needs depending on who is searching, the searcher's level of experience, and the content sought. Some databases are interdisciplinary, covering a large amount of content, while others are more focused and may be best suited to one field of research. Knowing where to strike the balance in choosing the appropriate databases to search can save time and energy when looking for specific content. A common method for gaining insight into the utility of research literature databases is to compare their effectiveness by performing sample searches to generate results and then analyzing the content retrieved.

This study examines eight research literature databases (AGRICOLA, AGRIS, BIOSIS, CAB Direct, FSTA, Google Scholar, Scopus, and Web of Science) and compares the search result retrieval for agronomy, sustainable diets, and meat science topics. In a prior study by the lead author (Ritchie et al. 2018), these same databases were evaluated for the scope of content included in these three topic areas. While the prior study used known-item searches to ascertain if specific citations were included in a database, this study attempts to reflect the searcher experience and determine if relevant content is retrieved using consistent search strategies. Standard information retrieval system evaluation metrics and statistical probability analysis of search result relevancy reveal the retrieval performance for each database. Interrater agreement analysis provides a richer interpretation of relevancy decisions used to generate the information retrieval system evaluation metrics.

Literature Review

Information Retrieval Evaluation Methods

The evaluation of information retrieval systems, especially for comparing the effectiveness of different systems to meet an information need, has a long history and established techniques. Common research methodology for evaluating information retrieval systems, including bibliographic databases, was pioneered by the Cranfield experiments starting in the 1950's and continues to be widely used in information retrieval evaluation efforts (Clough & Sanderson 2013). This methodology comprises selection of comparative strategies or systems, creation of a ranked list or pool of experimental results, determination of the relevancy of each item retrieved, measurement of the effectiveness of each strategy or system, and ranking the strategies or systems relative to each other. Figure 1 displays the research methodology for evaluation of information retrieval systems as adapted to bibliographic databases.

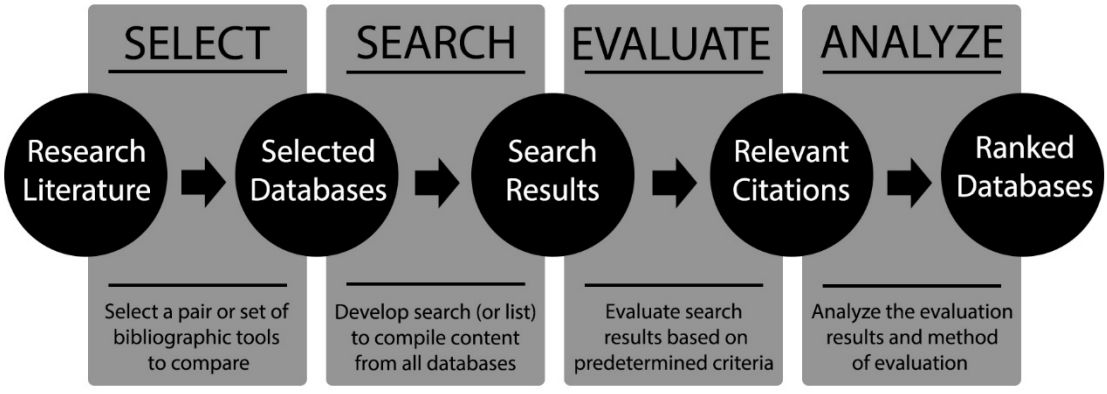

Figure 1. Evaluation of bibliographic databases as information retrieval systems

One feature of early (and continued) information retrieval work is the use of a test collection (a set of known documents with related queries and relevance judgments) to compare the performance of different systems (Clough & Sanderson 2013). In the 1990's, the U.S. National Institute of Standards and Technology (NIST), developed an ad-hoc method to build test collections that pooled search query results to create a representative set of documents (Sanderson 2010). As new search systems develop, research continues to explore methodologies to more efficiently and effectively build test collections using various strategies to compile or automatically harvest content for pooled results.

Evaluation metrics for effectiveness, particularly precision and recall, were also identified in the Cranfield experiments (Voorhees 2002; Clough & Sanderson 2013) and later included in early texts on information retrieval (van Rijsbergen 1979). These metrics of search retrieval have been further refined and adapted, and new metrics have developed as search systems and user preferences and needs have evolved (Clough & Sanderson 2013).

STEM Database Search Retrieval Comparison Studies

A review of database comparison studies reveals a range of information retrieval system methodologies and evaluation metrics in STEM and related disciplines, although all tend to conform to the accepted protocol established by the Cranfield experiments. Most studies alter the traditional methodologies, which utilize a test collection, in preference of developing ad-hoc document pools and using varied evaluation measures. Griffith et al (1986), queried five medical databases using both natural language and subject descriptor searches developed by eleven researcher and search-expert pairs. The researchers judged the relevancy and novelty of each set of results, and recall was calculated using the relevant documents retrieved for both types of search in all databases. Precision and recall were not significantly different, and the databases each retrieved a fair amount of unique content.

McCain et al (1987) searched five medical databases for eleven different topics and calculated recall, precision, and novelty ratios for each database on each topic. They teamed with eleven medical researchers to develop the search topics and strategies, and the researchers evaluated the relevancy of the pooled results. McCain et al determined that, while the methodology was labor intensive, it was most appropriate for comparison of databases with similar content. Stokes et al (2009) compared four health sciences bibliographic databases, with nursing students providing search topics and relevancy judgments. In addition to precision and recall metrics for effectiveness, uniqueness metrics were used to evaluate database efficiency and obtainability of cited content used to evaluate accessibility of relevant content. Joseph (2007) performed a simple search in eleven geology databases to evaluate results for relevancy and uniqueness. Search results were limited to a single publication year, and a retrospective search performed three years after the original search indicated significant new content had been added. A database overlap analysis was also a key feature of this study, with each database identified as having pertinent unique information on the search topic.

Ştirbu et al (2015) evaluated Google Scholar against Web of Science, FRANCIS, and GeoRef, for geographical content overlap and uniqueness, and also examined the changes in the results over time. A pool of results was created using a set of single keywords and limiting results to a four-year range. Sewell (2011) compared search result retrieval for CAB Abstracts across different database interfaces for veterinary literature. No difference between platforms was found for precision and recall across databases, revealing that platform is not likely to meaningfully impact search retrieval results. Walters (2009) compared Google Scholar to eleven other

databases for search retrieval of literature on the multidisciplinary topic of later-life migration. Each database was evaluated for precision and recall based on a core relevant set of 155 documents manually identified prior to searching. Google Scholar compared favorably with other databases, near the mid-top of all databases for precision and recall. This study's use of a core set of relevant documents is the closest to the test collection method.

Comparisons of Other Information Retrieval Systems

Studies of search result retrieval for search engines and other information retrieval systems also inform database comparison methodologies. Deka and Lahkar (2010) analyzed five different search engines (Google, Yahoo!, Live, Ask, and AOL) and their abilities to produce and then reproduce relevant and unique resources, focusing on the first ten results. Shafi and Rather (2005) compared precision and recall for five search engines (AltaVista, Google, HotBot, Scirus, and Bioweb), using a series of twenty searches on biotechnology topics. They used a relative recall calculation in their analysis to account for the large set of search results retrieved.

This study fills a gap in agricultural research literature database comparison, as no recent literature on search result retrieval for agricultural databases exists. Older studies that do exist were performed prior to the existence of Google Scholar and/or Scopus, and database content and functionality have changed greatly with the rapid evolution of online systems. Thus, findings from older studies may no longer be valid. As a result, recent studies for STEM disciplinary databases (some of which may also include agricultural content) informed the methodology for this study.

Relevance and interrater agreement

Common evaluation metrics in information retrieval comparison studies are based on a measurement of relevance. Relevance is a concept of interest to both developers and evaluators of information retrieval systems. It is non-binary (across individuals), context dependent, and subject to user interpretation (Buck 2017). An individual may judge something as relevant or irrelevant, but between individuals those judgments and understandings of what defines relevance often lie on a spectrum. Buck also notes that individuals have "different perspectives on what, at a particular time, is helpful information."

When relevance is determined by reviewers, results vary as reviewers agree or disagree on what is relevant. Interrater agreement measures the extent of agreement (McHugh 2012). Interrater agreement is used in education and psychology to develop rubrics and tests, in computer science to develop content analysis tools, in health sciences for clinical research and to train professionals, and in other areas where expert reviewers might agree or disagree. Voorhees (2002) discusses an assessor agreement method using overlap of relevant document sets by pairs (i.e., 1 & 2, 2 & 3, 1 & 3) to quantify the level of agreement by three reviewers. Lowe et al (2018) compared search results for eight multidisciplinary literature databases by differing types of search queries with a relevancy rubric shared by two evaluators or raters. Database overlap and relevancy were determined and contrasted across the raters to reflect differing user experiences.

Statistical Validation in Database Evaluation Studies

Studies in database evaluation often use statistical testing in order to determine any significance of their results; the chosen test relies on the nature of the data. Stokes et al (2009) used

Friedman's Test to identify if there were significant differences between the databases in the precision, novelty, and availability, as well as an odds ratio test to rate if the databases were, by their definitions, effective, efficient, or accessible. Deka and Lahkar (2010) used ANOVA and Tukey's HSD tests to evaluate if the difference between the mean number of relevant/stable hits from each database were significantly higher than any other databases. Sewell (2011) also used the ANOVA test to discover any significance in the different values of precision and recall between CAB Abstracts platforms.

Methods

Search construction

Predefined searches were conducted in spring 2017 in AGRICOLA (NAL Article Citation database only), AGRIS, BIOSIS, CAB Abstracts, FSTA, Google Scholar, Scopus, and Web of Science (see Table 1 for details; a description of each database can be found in Ritchie et al (2018)). No attempt to control for search retrieval variation across vendor platform was made, as previous research has shown no significant differences for precision and recall (Sewell 2011). Each database was searched for three different subject areas (agronomy, sustainable diets, and meat science). Search strings unique to each subject were created and aimed to retrieve at least one hundred results in each database for each topic. Search strings could include some use of Boolean operators, wildcard characters, and quotation marks (see Table 2). No additional filters or advanced search techniques were used, as these would have to be replicated across all databases and some do not include filters or limiters as an option. The search strings were adapted as needed for each database in accordance to their use of wildcard characters; for example, AGRICOLA required the use of a question mark, whereas other databases used an asterisk. The total number of results for each search was recorded, and then the results were sorted by date. Once the results were ordered by date, they were limited to those published

Table 1. Databases searched, publisher URL and platform searched

Database	URL	Platform/Vendor
AGRICOLA	https://agricola.nal.usda.gov/	National Agriculture Library
AGRIS	http://agris.fao.org/agris-search/index.do	Food and Agriculture Organization of the United Nations
BIOSIS	https://clarivate.com/products/web-of-science/databases/	Clarivate
CAB Abstracts	https://www.cabdirect.org/	CAB Direct
FSTA	https://www.ifis.org/fsta	EBSCO
Google Scholar	https://scholar.google.com/	Google
Scopus	https://www.scopus.com/	Elsevier
Web of Science	https://clarivate.com/products/web-of-science/	Clarivate

before 2016. Following the methodology used by Joseph (2007), a specific date range was identified to exclude the most recent citations to allow time for the ingestion and indexing of the materials by each database provider.

A few exceptions to searching were made depending upon the platform. Since Google Scholar only sorts by date for the current year, results were retrieved by using the date limiter to display the results of resources published during 2015. This exception could be made, as Google Scholar has been shown to generate more results than other databases (Ştirbu et al. 2015). AGRIS also did not sort results by date, so all those results were exported and then organized by date before evaluation.

The first one hundred citations in the results, sorted by date, and their accompanying metadata were then saved using the citation management tool Zotero. The first one hundred citations, displayed by date, were chosen, assuming researchers would not typically go past this number of results, and one hundred has previously been selected as a cut-off amount for developing ad hoc pools of results from searches (Sanderson 2010). Separate Collections (i.e., folders) were created for each database and subject combination in Zotero.

Table 2. Search strategies by topic

Agronomy	mycorrhiz* AND vegetabl*
Sustainable Diets	"sustainable diet" OR "sustainable diets" OR "econutrition" OR "sustainable food consumption" OR "sustainable nutrition" OR "ecological diet" OR "ecological diets" OR "resilient diet" OR "resilient diets" OR "biodiverse diet" OR "biodiverse diets"
Meat Science	"meat quality" AND retail*

Review of search results

Following the retrieval of the citations, all of the nearly 2,400 search result citations (some of the searches resulted in fewer than 100 results) were exported into spreadsheets to prepare the data for relevancy review. The review team consisted of three librarians with basic knowledge of agriculture research topics. Relevancy guidelines were provided by the lead researcher to the review team to guide decisions in addition to collaborators' own knowledge of the agricultural and health sciences (Table 3).

In order to manage the citations and review process, we used Rayyan (https://rayyan.qcri.org), a free tool developed for systematic review teams (Ouzzani et al. 2016). Rayyan allows users to upload structured citation data and abstracts from publishers and database providers, similar to many citation management programs, but also has a function to collaboratively review and mark citations as relevant or irrelevant (include/exclude) and provide reasoning for each determination. Determinations can be made in a blind mode so that no collaborator is influenced by the decisions of another. The review marking and masking functionality made Rayyan the best tool for relevancy reviewing.

Table 3. Relevancy determination guidelines for each topic

Agronomy: "Review each item as a researcher interested in the interactions between mycorrhizae - especially arbuscular mycorhizzal fungi - and vegetable crops. Interactions with cereal or other field/agronomic crops are acceptable. Look for impacts and interaction mechanisms. Items definitely not of interest are things about mushroom or ornamental plant culture."
Sustainable Diets: "Review each item as a researcher interested in diet and nutrition related to the environment or climate change. A few items have the word "sustainable/y" but they are not about environmental sustainability so exclude those items. Exclude articles on non-human diets and nutrition."
Meat Science: "Review each item as a researcher interested in meat quality as it impacts retail. These impacts can be on the display qualities, food safety traits, production systems, etc. as long as the article explicitly addresses both meat quality characteristics and retail. Retail can include retail display, marketing, consumer choices, etc. THE LINK TO RETAIL SHOULD BE CLEAR. For instance, articles that cover genomic traits that change meat quality characteristics commonly used to assess retail fitness would NOT be relevant. Articles that discuss customer response to different meat quality characteristics ARE relevant."

We uploaded spreadsheets of raw search result data, formatted to conform to the Rayyan citation metadata style, individually by database and topic. Once the data for each database was uploaded, reviewers used the title and abstract (if available) to determine the relevancy of each citation. The review of almost 2,400 citations took the review team several months to complete, as each citation was analyzed individually by each librarian. The individual reviewers normalized citations that appeared within multiple databases to ensure that the same article was consistently marked as relevant or irrelevant by the reviewer.

One of the drawbacks presented by Rayyan was the deduplication functionality. We did not deduplicate search results, so we could portray an authentic search experience, where any repeated record in a set of search results would be passed over by the typical researcher. However, Rayyan automatically identifies duplicate records in the same set of results and combines those records. In response, we retained spreadsheets of raw search results as references to control for duplicate search results removed in Rayyan.

Once the review was complete, the results were exported from Rayyan for analysis. Rayyan does not include the source database when exporting data. To efficiently retain the source, articles in Rayyan were filtered by database before exporting, creating one .csv file per database. In each file, rows represent articles. Reviewer data appeared as a string in one column; for example, RAYYAN-INCLUSION; {"Reviewer1"=>false, "Reviewer2"=>true, "Reviewer3"=>false}. True and false values were translated to zeros and ones in three columns, one per reviewer. A combined spreadsheet including each citation and three reviewer relevancy decisions (1 - relevant, 0 - irrelevant) was organized by topic area and database.

Data Analysis Methodology

In order to compare search result retrieval for each database, the team used a variety of tests to analyze different aspects of the results retrieved, both by database and by subject. Common metrics of precision, recall, and uniqueness formed the basis of the comparison. The team also investigated agreement among the three reviewers to determine whether overall relevancy scores

were subjective and to what degree. Finally, statistical testing highlighted any differences between databases considered significant and found similar patterns to the other relevancy tests.

Relevancy Calculation

Several techniques to rank or score relevancy are presented in the literature. In the most simple method, a single researcher reviews the results and determines if they are relevant based on a predefined scope and individual expertise (Joseph 2007; Sewell 2011). Some studies use a point system with specified criteria to compile a total score (Shafi & Rather 2005; Stokes et al. 2009; Lowe et al. 2018). Other studies use simple binary relevance judgments (yes/no) made by multiple reviewers to determine a total or averaged score for each set of search results (McCain et al. 1987; Deka & Lahkar 2010).

For this study, a binary value (1 or 0) was assigned to each citation by each reviewer; those three numbers were then added together to get a value (between 0 and 3) and then divided by the number of reviewers (3). This resulted in a value from 0 to 1 for each citation that was then summed to calculate our relevancy score for each database.

Thus,

Database Relevancy Score = Sum of the average relevancy score of each item retrieved

Precision Calculation

Precision is defined as the number of relevant results divided by the number of all returned results (Perry et al. 1955; van Rijsbergen 1979; Shafi & Rather 2005). For this study, a cutoff value of 100 was selected to emulate the likely behavior of a typical searcher. Selecting the "top" set of results based on assumptions about user behavior is a common technique for analysis of information retrieval systems and is referred to as precision at n (Craswell 2009). Although this technique does not determine the full precision of a complete set (or test set) of search results, it does allow for analysis of only those results likely to be reviewed by searchers (Craswell 2009). In a nuanced discussion of "measuring precision at a fixed ranking," Sanderson (2010) concludes that the "rank cutoff version of precision appears to be the better choice in most situations," especially if the user only wants to examine the first page(s) of results and with the caveat that other types of searches may require other measures.

$$\text{Precision (definition)} = \frac{\text{\# of relevant items retrieved}}{\text{\# items retrieved for search}}$$

$$\text{Precision at n} = \frac{\text{\# of relevant items retrieved}}{\text{\# items retrieved up to cutoff value}}$$

$$\text{Precision (this study)} = \frac{\text{Database Relevancy Score}}{\text{\# items retrieved up to cutoff value}}$$

Recall Calculation

Recall is often defined as the number of relevant items retrieved divided by the total number of relevant items possible to be retrieved (van Rijsbergen 1979). We made slight variations to the traditional recall definition for this study due to the large amount of relevant literature for our search topics and use of multiple raters.

First, since it would be very difficult to evaluate how many relevant items exist in all eight databases for each search, one can create a pool of relevant items comprised of the subset of relevant results retrieved from each of the eight databases. This methodology, referred to as relative recall, was developed by Clarke and Willett to assess search engine retrieval (another type of search retrieval with large numbers of relevant items) and is now commonly used for information retrieval analysis (Clarke and Willett 1997; Shafi & Rather 2005, Clough & Sanderson 2013). Clarke and Willett also account for the possible overlap of items retrieved between search result sets in their methodology, requiring an extra step to check that each item in the pool of all relevant items would actually be found in each search engine database and is not a duplicate of citations found in other databases. For this study, we have omitted the extra step, following a similar methodology used by Stokes et al (2009) and perhaps by Griffith et al (1986), although their methodology is not detailed nor referred to as relative recall. This simpler methodology assumes that all items in the relevant pool could be included in the content of each database searched, as all the items retrieved are likely within the scope of coverage for each. We also include duplicate citations as part of our pool although they were not considered relevant citations if they are duplicated within a database.

Secondly, because we are using an averaged relevancy score in place of numbers of relevant items retrieved, we used a sum of the Database Relevancy Scores across databases in each topic area to represent the total number of relevant results (i.e., the denominator of our recall calculation).

$$\text{Recall (definition)} = \frac{\text{\# of relevant items retrieved}}{\text{total \# of relevant items}}$$

$$\text{Relevant Recall} = \frac{\text{\# of relevant items retrieved}}{\text{\# relevant items pooled from all databases}}$$

$$\text{Recall (this study)} = \frac{\text{Database Relevancy Score}}{\text{sum of Database Relevancy Scores from all databases}}$$

Uniqueness Calculations

Uniqueness is defined as the number of citations retrieved from only a single database, i.e., not present in any of the other databases (Stokes et al. 2009; Deka & Lahkar 2010). For this study, the number of databases containing each citation was manually tabulated. Citations retrieved from only one database were counted as Unique. The average relevancy score for each unique citation in a database was summed to calculate the database's Unique and Relevant score.

Unique = # of items found only once in pooled results

Unique and Relevant = sum of the average relevancy scores of all unique items

Interrater Agreement Methodology

Interrater agreement is a measure of the degree to which raters make the same decision about an observation (McHugh 2012). In this study, interrater agreement allowed researchers to observe and interpret the variance in the judgments used for relevancy calculations and statistics. The raters in this case are agriculture and science librarians, and the data generated are binary values, indicating whether or not each article citation retrieved was relevant. We were less interested in determining the accuracy of each relevancy decision (i.e., did the raters correctly assign relevance according to a pre-determined standard), than we were in examining the consistency among raters, with the recognition that consistency of relevance decisions would normally vary among researchers using agricultural databases as well.

The most intuitive method of calculating interrater agreement is by percentage (Gwet 2014).

$$\text{Interrater agreement \%} = \frac{\text{\# of items all reviewers rated relevant + \# of items all reviewers rated irrelevant}}{\text{\# of items reviewed}}$$

McHugh (2012) describes the importance and methods of measuring interrater reliability and advises calculating both percent agreement and kappa to account for randomness among raters. To account for the agreement among pairs of the three reviewers, we calculated the pairwise percent agreement and Fleiss' kappa.

$$\text{Pairwise agreement} = \frac{\text{\# of items reviewers 1+2 both rated relevant + \# of items reviewers 1+2 both rated irrelevant}}{\text{\# of items reviewed by reviewers 1+2}}$$

Kappa is a commonly used index for interrater reliability and interrater agreement (Gisev et al. 2013). Fleiss kappa is a chance-corrected adaptation of the kappa index for nominal categories assessed by multiple raters (Gisev et al. 2013). Fleiss' kappa was calculated using the online tool ReCal3 (Freelon 2010). Gisev et al (2013) discuss variation in the interpretation of kappa values, noting Landis and Koch's (1977) six categories as the most comprehensive and widely cited and referencing Fleiss et al (2013) as a commonly used simplification. Both scales are used in the interpretation of the results.

Statistical Analysis Methodology

In order to perform the appropriate statistical analysis best suited for the study data, we consulted with a statistician. The statistician recommended using the raw data (not a comparison of means as in ANOVA tests) to create a generalized linear model, using the R language for statistical analysis to confirm if any databases were significantly more likely to produce a relevant result based on the reviewers' relevancy judgments. The combined spreadsheet of the relevancy decisions was reformatted for analysis in R, and the statistician provided the R code to produce the generalized linear model and the visual representation of the analysis. The generalized additive model specifically utilized in this study is a form of logistic regression used to analyze the relationships between variables to form a predictive model, which may not be linear (Crawley 2015). The variables analyzed using this test were the relevance scores, database, and subject, while accounting for the variability between the individual reviewers.

Results and Discussion

The data compiled from each of the twenty-four searches, subsequent refinement and selection, and relevancy determinations are summarized in Table 4.

Table 4. Summary of search and relevancy results by database and search topic.

Database / Subject	Total Hits	Total Hits (2015 & prior)	Total Retrieved	Database Relevancy Score
Agronomy				
AGRICOLA	108	106	100	86.00
AGRIS	217	216	100	77.33
BIOSIS	1,788	1,747	100	84.00
CAB	1,971	1,877	100	84.33
FSTA	132	118	100	67.67
Google Scholar	19,500	2,470*	100	78.33
Scopus	501	482	100	51.33
Web of Science	126	115	100	71.33
Totals	N/A	N/A	800	600.33

Database / Subject	Total Hits	Total Hits (2015 & prior)	Total Retrieved	Database Relevancy Score
Sustainable Diets				
AGRICOLA	82	59	59	40.00
AGRIS	113	107	100	55.33
BIOSIS	95	72	72	44.00
CAB	308	255	100	65.33
FSTA	123	83	83	57.67
Google Scholar	7,800	1,090*	100	62.33
Scopus	304	208	100	72.00
Web of Science	274	195	100	73.67
Totals	N/A	N/A	714	470.33

Database / Subject	Total Hits	Total Hits (2015 & prior)	Total Retrieved	Database Relevancy Score
Meat Science				
AGRICOLA	201	193	100	72.67
AGRIS	174	173	100	63.33
BIOSIS	208	197	100	67.00
CAB	827	783	100	69.00
FSTA	204	190	100	66.33
Google Scholar	14,400	1,180*	100	46.33
Scopus	203	188	100	65.67
Web of Science	308	278	100	68.67
Totals	N/A	N/A	800	519.00

*Due to the limitations of the Google Scholar platform, these values reflect the number of results limited to only the year 2015.

Precision

Precision represents the sum of the average relevancy scores of the set of all items retrieved for each database by topic at the cutoff value of 100.

$$\text{Precision (this study)} = \frac{\text{Database Relevancy Score}}{\text{\# items retrieved up to cutoff value}}$$

Table 5. Summary of precision for each database by topic. Above average precision is indicated in bold.

	Agronomy	Sustainable Diets	Meat Science	Database Average
AGRICOLA	**0.860**	**0.678**	**0.727**	0.755
AGRIS	**0.773**	0.553	0.633	0.653
BIOSIS	**0.840**	0.611	**0.670**	0.707
CAB	**0.843**	0.653	**0.690**	0.729
FSTA	0.677	**0.695**	**0.663**	0.678
Google Scholar	**0.783**	0.623	0.463	0.623
Scopus	0.513	**0.720**	**0.657**	0.630
Web of Science	0.713	**0.737**	**0.687**	0.712
Topic Average	0.750	0.659	0.649	

Since we retrieved and rated 100 results for each database for each topic (with a few exceptions), precision is frequently the percentage of the averaged relevancy score for each database by topic. The agronomy topic had the highest average precision across all databases for any topic. The majority of databases had above or near average precision for the topic with the exception of FSTA and Scopus. For FSTA, this is explained by agronomy being outside of the topical scope. For Scopus, non-relevant results tended to deal with genomics or non-vegetable/field crops such as greenhouse ornamentals. For the sustainable diets topic, the broad-scope multidisciplinary databases, Scopus and Web of Science, as well as FSTA, the food and nutrition topic specific database, retrieved the most relevant search results. Precision for this topic across all databases clustered around the average with no large variations. Meat sciences search retrieval results were the least precise as averaged across all databases, with Google Scholar search results rated much less relevant than the other databases. Non-relevant Google Scholar results lacked the retail aspect of the search topic. AGRICOLA was the only database with above-average precision of search retrieval for all topics. As AGRICOLA is focused on comprehensively collecting agricultural research, it is expected that items retrieved were relevant across agricultural topics.

Recall

Recall represents the sum of the average relevancy scores divided by the total number of relevant items retrieved across databases for each topic.

$$\text{Recall (this study)} = \frac{\text{Database Relevancy Score}}{\text{sum of Database Relevancy Scores from all databases}}$$

Table 6. Summary of recall for each database by topic. Above average recall is indicated in bold.

	Agronomy	Sustainable Diets	Meat Science	Database Average
AGRICOLA	**0.143**	0.085	**0.140**	0.123
AGRIS	**0.129**	0.118	0.122	0.123
BIOSIS	**0.140**	0.094	**0.129**	0.121
CAB	**0.141**	**0.139**	**0.133**	0.137
FSTA	0.113	0.123	**0.128**	0.121
Google Scholar	**0.131**	**0.133**	0.089	0.117
Scopus	0.086	**0.153**	**0.127**	0.122
Web of Science	0.119	**0.157**	**0.132**	0.136
Topic Average	0.125	0.125	0.125	

By using a modified relative recall method, we decided we were more interested in determining what was relevant in the set of retrieved items, and not determining every item which could have been retrieved. Random spot checks of our data reveal that items counted as unique within the search results could be found in the contents of other databases, but not in the top set of results. However, the focus of this study was the search results, not the database content coverage, which has previously been explored using methods more suited to examining database content (Ritchie et al. 2018). We are not designing or trying to improve database performance against a perfect set of results, rather we are evaluating the actual performance of one database against another.

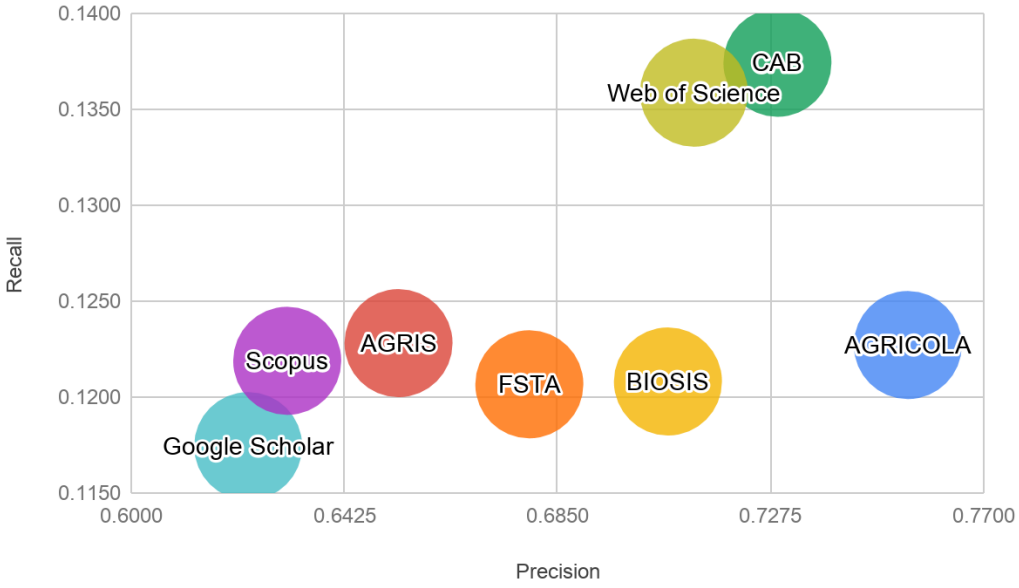

Figure 2. Databases plotted by average precision and recall for all topics

CAB and Web of Science both had well above the average recall across topics, with CAB performing above average for all topics and Web of Science with above average recall for the multidisciplinary sustainable diets topic. Conversely, AGRICOLA had a lower average recall score, which was affected by its low recall score for the sustainable diets topic (only 59 citations retrieved for the search). It is possible a complete retrieval for the sustainable diets topic would have resulted in an average recall similar to CAB or Web of Science. BIOSIS and FSTA also had lower than target result retrieval for the sustainable diets topic, which could be expected for a biological sciences database like BIOSIS, but was surprising for the food and nutrition focused FSTA. However, almost all of the lesser number of FSTA results were relevant, so the recall in this topic was near average for FSTA. Figure 2 provides a visual comparison of precision and recall for each database to illustrate the data summarized in Tables 5 and 6.

Uniqueness

Uniqueness is the sum of all citations only retrieved from a single database. Unique and relevant citations are those that are both rated as relevant by reviewers and only retrieved from a single database.

Table 7. Summary of uniqueness for each database by topic.

	Agronomy		Sustainable Diets		Meat Science		Average	
Databases	**Unique Items**	**Averaged Relevancy Score for Unique Items**	**Unique Items**	**Averaged Relevancy Score for Unique Items**	**Unique Items**	**Averaged Relevancy Score for Unique Items**	**Unique Items**	**Averaged Relevancy Score for Unique Items**
AGRICOLA	16	14.67	4	2.33	15	10.00	11.67	9.00
AGRIS	25	18.00	35	20.00	19	10.67	26.33	16.22
BIOSIS	68	57.33	23	9.67	13	10.00	34.67	25.67
CAB	51	41.33	36	21.00	65	46.67	50.67	36.33
FSTA	78	48.00	20	12.67	22	12.67	40.00	24.44
Google Scholar	73	54.67	62	36.33	92	38.33	75.67	43.11
Scopus	69	26.67	13	8.00	15	10.67	32.33	15.11
Web of Science	43	25.00	13	9.67	40	23.67	32.00	19.44
Total	423	285.67	206	119.67	281	162.67	303.33	189.33

Uniqueness of search retrieval content can be important for some searches, especially those for which an exhaustive recall of items is important, such as to support a systematic review or a patent application. Additionally, a key factor that influences collection decisions is whether an information source finds content unavailable/unretrievable through other sources. Databases that retrieve items that are both unique and relevant will likely be of use to many searchers and can help to expand a library's ability to provide content from varied sources.

CAB and Google Scholar on average retrieved the most unique, and unique and relevant content. However, almost half of the unique Google Scholar citations were not relevant. This result is very similar to that of Ştirbu et al (2015), where only half of the Google Scholar citations retrieved for geographic literature were considered relevant to the topic. By count, non-relevant Google Scholar citations numbered more than twice the average of all unique, but not relevant items (blue portion of Figure 3) for all databases. By percentage, Google Scholar returns the second most non-relevant unique citations, but within the same range as other databases. Almost all the unique items retrieved from AGRICOLA were rated as relevant, and although this was the smallest total number of unique items, by percentage, unique results from AGRICOLA remain the most relevant. BIOSIS and CAB also return the mostly relevant unique results by percentage. These three databases could be a better option for the searcher that does not wish to spend extra effort and time to scrutinize results.

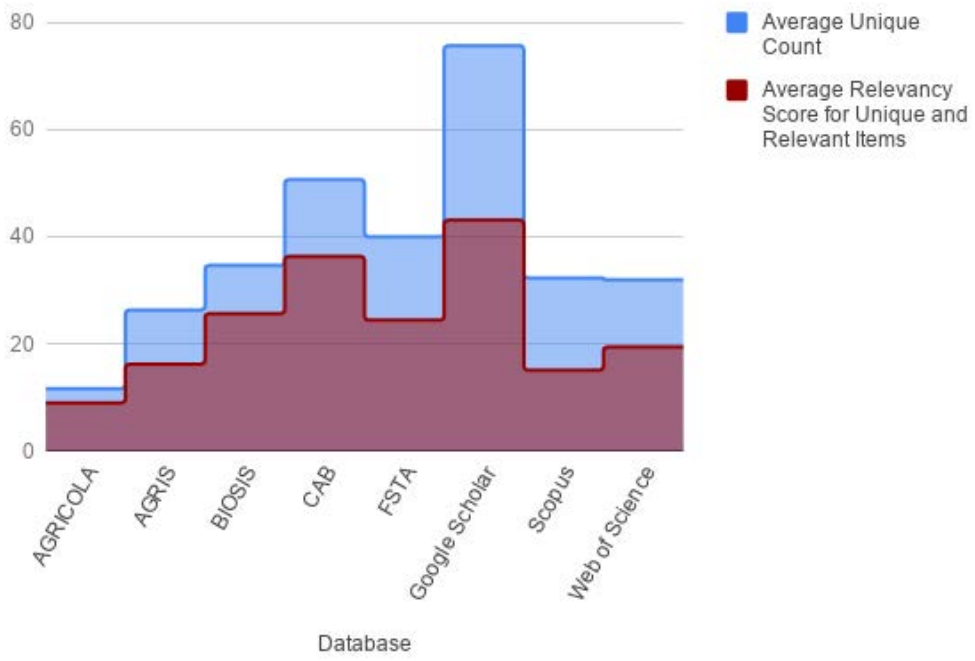

Figure 3. Average Uniqueness of Citations

The uniqueness graph displays the average count of unique citations across topics for each database overlaid with the averaged relevancy score of unique and relevant citations averaged for all topics.

Interrater Agreement

Interrater agreement is a measure of the degree of consistency among reviewers.

$$\text{Interrater agreement \%} = \frac{\text{\# of items all reviewers rated relevant + \# of items all reviewers rated irrelevant}}{\text{\# of items reviewed}}$$

Our subjective determination of relevancy and varying domain expertise led to different relevancy determinations for articles (see Figure 4). This is realistic. Researchers on a spectrum of expertise might agree or disagree on what is useful, and the same researcher might deem the same item relevant or not at different times in their research process.

Relevance judgments were based solely on content guidelines for each topic and, thus, did not account for

- whether or not the abstract or a link to the article was present.
- language of the title and abstract being reviewed.
- type of resource being reviewed (e.g., reference or primary research sources).
- a common process for completing reviews (e.g., length of time or number of reviews in each session).

Despite subjectivity and diversity in methods for performing the reviews, interrater agreement data demonstrate a moderate to substantial agreement by the Landis and Koch interpretation, or fair to good agreement by the Fleiss et al (2013) interpretation according to Fleiss' kappa interpretations discussed in Gisev et al (2013), see Table 8. Google Scholar had the lowest agreement for all topics followed by CAB, which also ranked below average on all topics. This could be related to uniqueness, as these two databases also ranked highest in that metric. Google Scholar is also less consistent with the availability and quality of summary information (i.e., abstract) which was used to determine relevance. BIOSIS had above average agreement on all topics. Some of our topics were out of the scope of this biological sciences database so the percentage of agreement is influenced by consensus on irrelevant articles.

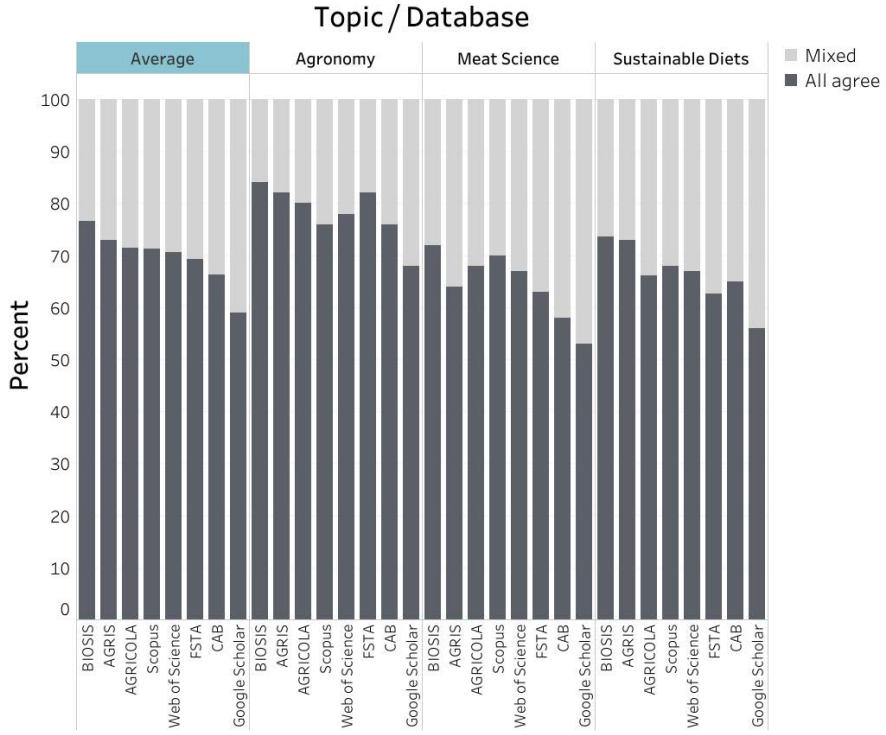

Figure 4. Percent of documents in which all three reviewers agreed on relevancy, organized by topic and database. In the sustainable diets topic, some databases did not retrieve 100 results.

Of the topics searched and reviewed, agronomy had the highest agreement for each database. Agronomy was the first topic reviewed, and reviewers indicated that it was either the most

familiar topic or that it was the easiest topic to identify the irrelevant citations (i.e., wrong type of mycorrhizae). Reviews may have been influenced by reviewer engagement with a topic through experience and the review process. Rayyan provides a summary of the length of time and number of sessions for each reviewer, both of which might influence a reviewer's judgment.

Table 8. Average pairwise agreement and Fleiss' kappa by topic and database.

	Total articles	Articles Reviewers Included			Articles Reviewers Excluded			Average Pairwise Agreement	Fleiss' kappa
		R1+R2	R2+R3	R1+R3	R1+R2	R2+R3	R1+R3		
Agronomy	800	593	518	516	132	165	128	85.5	0.61**
Sustainable Diets	714	455	351	364	146	197	147	77.47	0.49*
Meat Science	800	464	415	395	212	164	184	76.42	0.48*
AGRICOLA	260	194	162	168	32	48	29	81.47	0.48*
AGRIS	300	189	161	157	78	80	73	82	0.60*
BIOSIS	273	189	166	167	59	59	50	84.56	0.62**
CAB	300	216	171	168	41	53	49	77.56	0.43*
FSTA	284	191	151	148	61	66	62	79.98	0.54*
Google Scholar	300	157	146	135	73	72	71	72.67	0.42*
Scopus	300	177	152	153	88	84	76	81.11	0.59*
Web of Science	300	199	175	179	58	64	49	80.44	0.52*

Note: The sustainable diets searches retrieved less than 100 results from AGRICOLA, BIOSIS, and FSTA.
All Fleiss' kappa results in this table are in the fair to good range on Fleiss' scale (Fleiss et al. 2013; Gisev et al. 2013). Using kappa interpretations by Landis and Koch (1977), * indicates moderate agreement and ** indicates substantial agreement (Gisev et al. 2013).

Statistical Analysis

The analysis shows the results of the generalized additive model, displaying the estimated probability values, or the chance from 0 to 1 that the database will retrieve a relevant article, as well as the 95% confidence intervals while accounting for the differences between reviewers as a variable (see Figure 5). Where there is overlap, the databases are similar in their likelihood of producing relevant articles; no overlap between database confidence intervals indicates a significant difference in this likelihood.

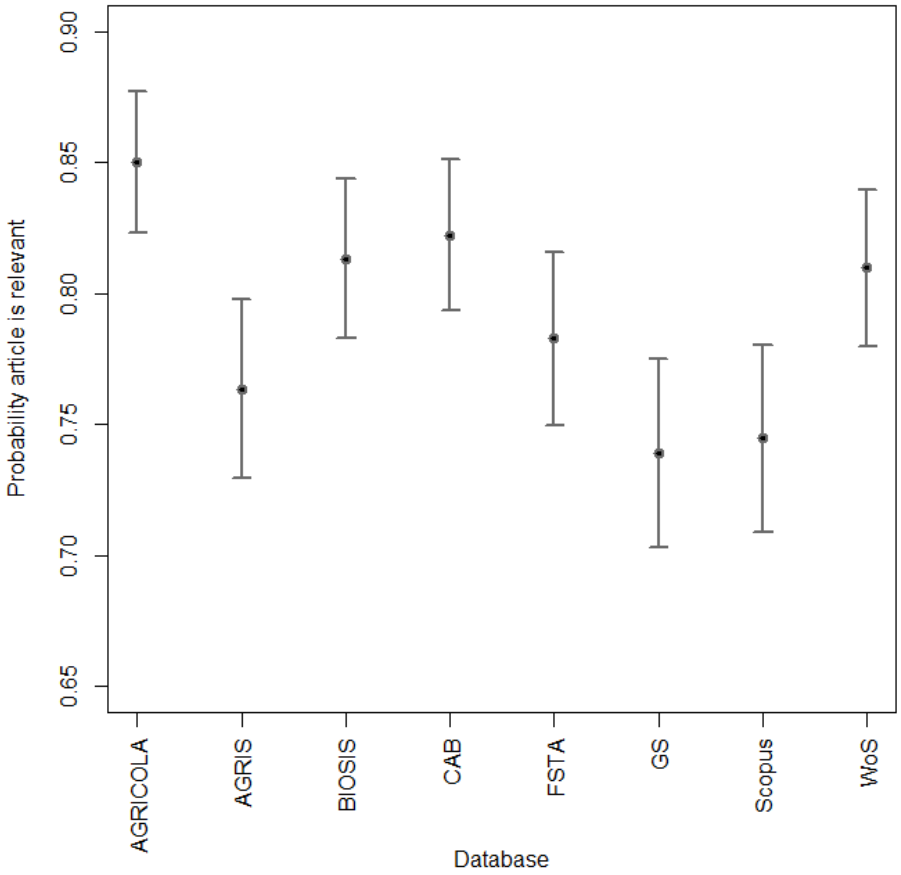

Figure 5. The dots are the estimated probability values of the database producing a relevant article. The lines are the 95% confidence interval.

Overall Probability of Relevance

Using R to produce a summary of the data from the generalized additive model, the databases are ranked from highest to lowest in the probability of finding a relevant article, irrespective of subject, utilizing this study's data pool: AGRICOLA, CAB, BIOSIS, Web of Science, FSTA, AGRIS, Scopus, and Google Scholar. The analysis shows there is some significance between certain databases; for example BIOSIS and Web of Science have very close estimated probabilities, while the confidence intervals for AGRICOLA and Google Scholar do not overlap at all and therefore are significantly different in their probabilities of finding relevant articles.

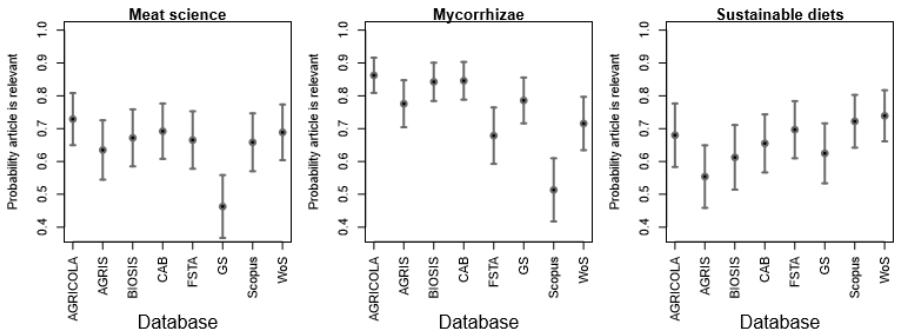

Figure 6. The dots are the probability values of the databases producing a relevant article, organized by subject. The lines are the 95% confidence interval.

Probability of Relevance by Subject

The analysis by subject (Figure 6) shows that while some databases, such as AGRICOLA, consistently had a comparatively higher likelihood of producing relevant results, other databases were stronger in specific disciplinary areas. This makes sense due to the nature of the different databases and the types of content found within them. For example, Scopus, a broad, multidisciplinary database, did not have a high estimated probability of finding an article regarding mycorrhizae, but had a much higher estimated probability of finding an article regarding sustainable diets. This is a database without a specific focus, and so it may contain more relevant content in some areas more than others. A more subject specific database like AGRICOLA had a higher probability due to its more concentrated content.

When discussing odds ratios, Stokes et al (2009) warn that results where confidence intervals overlap may not be consistently reproduced; similar advice can be applied here. The results show that there is some consistency in databases across all subjects, and some subject areas, particularly agronomy, had higher probabilities of having relevant results than others. This could be attributed to the searched topic and nature of the literature within these certain fields; broader keywords could be interpreted and utilized differently across multiple disciplines and mislead researchers into thinking the number of results is equal to number of relevant results. For instance, the term mycorrhizae has a much more specific definition than sustainability. If looking to compare databases within a subject area, the statistical analysis could reveal significant differences in relevant content.

The blind-review methodology assisted in maintaining independence of observations between reviewers, but the research team did individually normalize decisions on articles duplicated across databases. While this violates the independence of observations, it was deemed necessary to ensure a fair analysis of each database.

Search Limitations

The study design was not constructed to evaluate the relevance algorithms of each database, simply the content retrieval. Databases are continuously updating not only their content but also their platforms, making the analysis of relevance algorithms time-sensitive (Bethel & Rogers 2014). Due to possible variations in relevance algorithms, this study decided to rely on date ranges to get a more constant measurement of the content contained in each one. Joseph (2007), however, found that even when limited to a specific date range, number of results still varied between identical searches performed in different years due to the addition or deduplication of materials within the evaluated databases. Additionally, Google Scholar literature was limited to just one year for this study, which may have affected comparability to the other databases.

Although this study attempted to get more constant results by using date ranges rather than relevance algorithms, this may not be completely successful across the platforms. Google Scholar, especially, showed more variation in results over time than other examined databases in the evaluation done by Ştirbu et al (2015). Google Scholar itself is a unique case, as searches performed within Google Scholar are not easy to reproduce because the platform alters the algorithm frequently and has limitations on filtering by date (Bramer et al. 2016). Search strategies in other, more conventional databases are harder to convert into optimized Google Scholar searches, which other researchers have produced by using different Boolean characters and relocating quotation marks (e.g., Bramer et al. 2016). This study did not attempt to translate

the search, as our target search was that of an intermediate searcher which may have led to less relevant results.

Conclusion

Overall, AGRICOLA search results are the most precise or relevant, and CAB and Web of Science search results recall more of the total relevant citations than other databases. Each of these three databases comprise highly curated content for the agricultural disciplines. Researchers are likely to meet information needs using these databases dependent on how comprehensively they would like to cover a topic. FSTA and BIOSIS have slightly less overall relevance, which might be explained by some of the topics being outside of their stated content scope. AGRIS would have been expected to have higher precision, as it is focused on agricultural content, but many duplicate records rated as irrelevant might have reduced overall precision. Scopus and Google Scholar, as broader, multidisciplinary databases with more permissive guidelines for content inclusion, performed poorly in terms of precision and recall, as compared to the other databases.

Reviewer agreement is a metric used to elucidate results of comparison studies involving multiple reviewers. The goal of reviewing articles in this study was not to eliminate bias, create a gold standard, or develop an assessment tool. Subjectivity was expected and apparent in the interrater agreement calculations. Despite this variance, interrater agreement calculations demonstrate moderate reliability among reviewers. These results provide insights when evaluating information retrieval systems, as the analysis of agreement captures the subjectivity of relevance determinations and their impact upon other information retrieval measures, contributing to the interpretation of the information retrieval measures in the context of system users' differences. Overall, interrater analysis provides a richer understanding of search result retrieval analysis and the retrieval measures themselves. To achieve higher agreement while preserving individual judgments, future studies comparing databases by multiple reviewers judging results may want to control for some of the considerations in our relevance judgments discussion: expertise, language, metadata, type of resource, time, and reviewers per session.

The statistical tests mostly supported the results of the raw data analysis, finding similar patterns across the databases and subject areas. The statistical comparisons of the probabilities of finding relevant articles in certain databases were not significant, demonstrating they would be about equally effective in producing a relevant article within their results. Others showed a significant varying in degree of probability, suggesting that certain databases may be likelier to produce relevant results, and which could be affected by subject area of the search and the concentration (if any) of the database. These results, combined with the results of the precision and recall calculations, can help to create a clearer picture of which databases can meet specific research needs. While the statistical significance may not be drastic or present between many of the databases, a perceived lack of relevant content by researchers can dissuade them from utilizing a database; a database containing a higher percentage of relevant results for a subject area of interest would be a good place for researchers to start their search and for libraries to invest their resources. While Sewell (2011) did not find any statistically significant advantages to certain platforms to access the same resource, the article acknowledges the importance of user perspectives in regard to platform specific functionalities and usability. Stokes et al (2009) note that relevancy does not consider all user needs, which would require further analysis, but it is necessary for testing a databases "effectiveness".

This study relied not only on statistical testing, but multiple factors to evaluate each database. The results of this study indicate which databases can best help researchers meet a range of information needs within agricultural disciplines and select the most appropriate bibliographic database based on their project requirements. Searchers requiring a few highly relevant results, such as to support a student research paper, should use a database with high precision (AGRICOLA). Searchers requiring comprehensive information about a topic should consider a database with high recall (CAB or Web of Science). Searchers performing an exhaustive review should include sources of unique and relevant content (Google Scholar or CAB). This study identifies appropriate databases tools to meet these needs and can save a searchers time, effort and frustration.

Additionally, this study could be used to manage electronic resources within institutions. Librarians can use the study results to determine which bibliographic databases meet the requirements of their patrons. It is important for librarians to identify those databases with the best chance of producing relevant results for users to ensure patron success, which can potentially drive acquisitions decisions. Additionally, librarians can use this study to support recommendations to a range of stakeholders regarding the value of a particular agricultural database, especially if faced with budgetary decisions.

Acknowledgements

The authors would like to thank Dr. Trevor Hefley for consulting and performing the statistical analysis.

References

Bethel, A. & Rogers, M. 2014. A checklist to assess database-hosting platforms for designing and running searches for systematic reviews. *Health Information & Libraries Journal* 31(1): 43–53. DOI: 10.1111/hir.12054.

Bramer, W.M., Giustini, D. & Kramer, B.M.R. 2016. Comparing the coverage, recall, and precision of searches for 120 systematic reviews in Embase, MEDLINE, and Google Scholar: A prospective study. *Systematic Reviews* 5(1): 39. DOI: 10.1186/s13643-016-0215-7.

Buck, W. 2017. Precision and recall: An ontological perspective. *Canadian Journal of Information & Library Sciences*, 41(1/2): 42–51. https://muse.jhu.edu/article/666448.

Clarke, S.J. & Willett, P. 1997. Estimating the recall performance of web search engines. *Aslib Proceedings* 49(7): 184–189. DOI: 10.1108/eb051463.

Clough, P. & Sanderson, M. 2013. Evaluating the performance of information retrieval systems using test collections. *Information Research* 18(2). Paper 582. http://www.informationr.net/ir/18-2/paper582.html.

Craswell, N. 2009. Precision at n. In: Liu L, Özsu MT, editors. *Encyclopedia of Database Systems*. Boston, MA: Springer US. p. 2127–2128. DOI: 10.1007/978-0-387-39940-9_484.

Crawley, M.J. 2015. *Statistics: An Introduction Using R*. 2nd ed. Chichester (UK): John Wiley & Sons. DOI: 10.1002/9781119941750.

Deka, S.K. & Lahkar, N. 2010. Performance evaluation and comparison of the five most used search engines in retrieving web resources. *Online Information Review* 34(5): 757–771. DOI: 10.1108/14684521011084609.

Freelon, D.G. 2010. ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science* 5(1): 20-33. http://www.ijis.net/ijis5_1/ijis5_1_freelon.pdf.

Fleiss, J.L., Levin, B. & Paik, M.C. 2013. *Statistical methods for rates and proportions*. John Wiley & Sons. DOI: 10.1002/0471445428.

Gisev, N., Bell, J.S. & Chen, T.F. 2013. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy* 9(3): 330–338. DOI: 10.1016/j.sapharm.2012.04.004.

Griffith, B.C., White, H.D., Drott, M.C. & Saye, J.D. 1986. Tests of methods for evaluating bibliographic databases: An analysis of the National Library of Medicine's handling of literatures in the medical behavioral sciences. *Journal of the American Society for Information Science* 37(4): 261–270. DOI: 10.1002/(SICI)1097-4571(198607)37:43.0.CO;2-6.

Gwet, K.L. 2 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters* (4th edition). Gaithersburg, MD: Advanced Analytics, LLC. http://www.agreestat.com/book4/.

Joseph, L.E. 2007. Comparison of retrieval performance of eleven online indexes containing information related to quaternary research, an interdisciplinary science. *Reference & User Services Quarterly* 47(1): 56–65. DOI: 10.5860/rusq.47n1.56.

Landis, J.R. & Koch, G.G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159–174. DOI: 10.2307/2529310.

Lowe, M.S., Maxson, B.K., Stone, S.M., Miller, W., Snajdr, E. & Hanna, K. 2018. The Boolean is dead, long live the Boolean! Natural language versus Boolean searching in introductory undergraduate instruction. *College and Research Libraries* 79(4): 517–534. DOI: 10.5860/crl.79.4.517.

McCain, K.W., White, H.D. & Griffith, B.C. 1987. Comparing retrieval performance in online data bases. *Information Processing & Management* 23(6): 539–553. DOI: 10.1016/0306-4573(87)90058-6.

McHugh, M.L. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica* 22(3): 276-282. DOI: 10.11613/BM.2012.031.

Ouzzani, M., Hammady, H., Fedorowicz, Z. & Elmagarmid, A. 2016. Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews* 5: 210. DOI: 10.1186/s13643-016-0384-4.

Perry, J.W., Kent, A. & Berry, M.M. 1955. Machine literature searching X. Machine language; Factors underlying its design and development. *American Documentation* 6(4): 242–254. DOI: 10.1002/asi.5090060411.

Ritchie, S.M., Young, L.M. & Sigman, J. 2018. A comparison of selected bibliographic database subject overlap for agricultural information. *Issues in Science & Technology Librarianship* 89. DOI: 10.5062/F49Z9340.

Sanderson, M. 2010. Test collection based evaluation of information retrieval systems. *FNT in Information Retrieval* 4(4): 247–375. http://marksanderson.org/publications/my_papers/FnTIR.pdf.

Sewell, R.R. 2011. Comparing four CAB Abstracts platforms from a Veterinary Medicine perspective. *Journal of Electronic Resources in Medical Libraries* 8(2): 134–149. DOI: 10.1080/15424065.2011.576608.

Shafi, S.M. & Rather, R.A. 2005. Precision and recall of five search engines for retrieval of scholarly information in the field of Biotechnology. *Webolog* 2(2). http://www.webology.org/2005/v2n2/a12.html.

Ştirbu, S., Thirion, P., Schmitz, S., Haesbroeck, G. & Greco, N. 2015. The utility of Google Scholar when searching geographical literature: Comparison with three commercial bibliographic databases. *Journal of Academic Librarianship* 41(3): 322–329. DOI: 10.1016/j.acalib.2015.02.013.

Stokes, P., Foster, A. & Urquhart, C. 2009. Beyond relevance and recall: Testing new user-centred measures of database performance. *Health Information & Libraries Journal* 26(3): 220–231. DOI: 10.1111/j.1471-1842.2008.00822.x.

van Rijsbergen, C.J. 1979. *Information Retrieval*. 2nd ed. London: Butterworths. http://www.dcs.gla.ac.uk/Keith/Preface.html.

Voorhees, E.M. 2002. The philosophy of information retrieval evaluation. In: Peters C, Braschler M, Gonzalo J, Kluck M, editors. *Evaluation of Cross-Language Information Retrieval Systems*. CLEF 2001. Vol. 2406. Berlin: Springer. (Lecture Notes in Computer Science). p. 355–370. DOI: 10.1007/3-540-45691-0_34.

Walters, W.H. 2009. Google Scholar search performance: Comparative recall and precision. *portal: Libraries & the Academy* 9(1): 5–24. DOI: 10.1353/pla.0.0034.