

Matthew Stephen Treskon

Digital Services Librarian—DigiTop | National Agricultural Library
mtreskon@nal.usda.gov | (Ph) 301-504-5204 | (F) 301-504-6409

Utilizing Citation Analysis to Assess the Coverage of Abstracting & Indexing Databases

Abstract

By utilizing citation analysis at the article level, this study identified the demonstrated patron need of Agricultural Research Service (ARS) researchers of the United States Department of Agriculture (USDA), and then used this metric to evaluate the coverage of four Abstracting & Indexing services (Scopus, Agricola, Biosis, and CAB), as well as DigiTop's full-text journals. It is believed that this study will provide useful findings for collection development and patron instruction at the National Agricultural Library (NAL). This study also seeks to illustrate the feasibility and utility of citation analysis at the article level for the evaluation of library service.

Introduction

A&I services are designed to enable/facilitate the discovery, evaluation and access of material. Although it used to be the case that A&I services were the primary, if not the only, way to enable patron access to some types of materials, this is often not the case; web-based search engines, such as Google, are now the predominate method for many types of search strategies. However, it has been argued in the library community and elsewhere, that despite the

pervasiveness of these tools, the structure and descriptiveness of the A&I databases still make these services as useful today as they were yesterday. Despite the library community continued commitment to these services, these services are only occasionally evaluated.

This study evaluated four Abstracting & Indexing databases according to the needs of USDA-ARS researchers. Ideally, a full evaluation of the utility of these databases would include the four major components of an A&I database: coverage, quality of citations, the search algorithm, and the user interface. However, due to time constraints, the author focused on coverage, contending that it is worth independent analysis because it is the foundation upon which the other components are built. Regardless of how well constructed the other components may be, they can only be as good as the content they are designed to access. A future study might combine a bibliometric analysis similar to the one detailed in this paper, with an end-user study designed to evaluate the descriptive utility of the citations, effectiveness of the search algorithm, and usability of the interface.

Researcher need, for the purposes of this study, was identified by extracting a representative sample of works cited from ARS publications. Although measures such as journal title coverage and access records (such as search logs) were considered, the author decided that this method provided the most representative measure of researcher need, given time and labor constraints.

Although conducted within the specific context of the National Agricultural Library, and to shed light on the relative utility of the various A&I services for USDA-ARS researcher need, it is hoped that this study will not be read without consideration to other contexts. It is this author's belief that the methodology and results of this study are applicable to the greater research library community.

The Context

The study has been designed to evaluate services at the National Agricultural Library according to the needs one of its main constituent groups, ARS researchers. Although it is believed that there are more similarities than differences between the National Agricultural Library and other research libraries, there are important and unique characteristics of the library and its patrons that need explanation.

First, the library serves both USDA and the agricultural community at large. Although a complete evaluation of need would consider both service to both constituent groups, this study focused on the needs of USDA in general, and ARS researchers in particular. It is believed that this study's analysis of researcher publications (the end result of research) and their works cited (their literary origins) provides an extremely representative measure of this need.

Another important characteristic of the library is that one of the primary constituent groups, USDA-ARS researchers,⁴ are highly knowledgeable of their respective fields. Studies have indicated that the level of research and domain-specific expertise a library patron is correlated with the extent of their knowledge of information resources.¹ For the purposes of this study, it is assumed that USDA-ARS researchers' high degree of expertise translates into a high degree of information literacy, and that this knowledge base mitigates the effect of differential rates of access to material on the product of their research, including their works cited.

Finally, the National Agricultural Library is distinguished from many other academic research libraries by its focus on agriculture and related fields. As a result, it is easier for the library to tailor its collection and services (including A&I services) than libraries whose collection must cover a wide-range of disciplines. Similarly, this focus also means that it is significantly easier to evaluate the library's collection and services by subject analysis.

However, this is not to imply that the difficult questions of scope do not exist. The questions, *what is agriculture*, and *what are the relevant related fields* for collection development purposes, are questions that the library has to continually address. Furthermore, the development of multidisciplinary research has led to the blurring of disciplinary boundaries, and the subject-based library has had to contend with that which, by nature, defies traditional classification.

Although these issues of scope need to be addressed, they do not necessarily need to be addressed by an evaluation of coverage. Since the methodology of this study does not require a determination of subject, the issues of classification do not need to be considered.

Literature Review of A&I services

Even though there are now many avenues to literature discovery, and new technologies such as federated searching and digital libraries merit our professional interest, A&I services are still being used and should continue to be discussed. If and how much they are being used, is often unknown. This study is a limited attempt to address that question, focusing on one aspect of A&I services, their coverage, in one particularly context, ARS researchers.

Bibliometric studies of patron need can be divided into reader-based or author-based, and can be conducted at the journal title (or source) level, or the article (item) level. Reader-based metrics are typically based on download/database access statistics whereas author-based metrics are typically based on patron publications.

It could be argued that download/access statistics would be a better measure of need, since they include material that the researcher found interesting/educational but not necessarily relevant for citation. Within a limited domain, such as a digital library, where all access data is of the same format and can therefore be easily aggregated, reader-based metrics can be conducted. Bollen et al² designed an alternative to the ISI journal impact factor, based on a reader-generated network derived from social network status metrics .

However, this limited domain does not represent today's research library. Materials needed for research are in multiple formats, as is the data of their usage. Subsequently, a comprehensive reader-based metric of patron need at a given research library is not feasible. Although the development of the COUNTER Code of Practice and the SUSHI protocol have helped standardize usage data reporting and facilitate data collection, issues with uniform reporting still exist. Furthermore, the definition of what constitutes a statistical count, (a request, search, or session), has never been etched in stone, and this uncertainty has only been

exasperated by the mashing/integration of services available in the Web 2.0, including federated searching.³

At some point in the future, it is possible that usage statistics of all formats will be similar enough to allow a comprehensive reader-based usage metric. However, that is not currently the case, and as a result, a review of the literature suggests that reader-based metrics have not been applied to the study of A&I services due to this complexity.

On the other hand, author-based metrics, or citation analysis studies, are feasible and have been so for some time. These studies have been designed to aid the institution in quantifying research 'output,' determine the relative influence of a journal, study social networking research patterns, aid collection development, and evaluate the coverage of A&I services.

Author-based and reader-based metrics can be applied to the evaluation of A&I services at either the subject, journal title, or article level. These evaluative studies have typically been designed for either the selection or the development of these databases. Evaluative studies for coverage of A&I services for the purposes of selection have primarily utilized journal-level analysis, and only rarely article-level analysis. Subject disciplines have ranged from library and information science (Jasco 1998)⁴ and (LaBorie 1984)⁵, medicine,⁶ religion,⁷ and agriculture⁸ to name a few. Coverage for multidisciplinary fields for which the application A&I services and other information services has proven particularly troublesome, such as neuroscience, have also been conducted⁹.

Unfortunately, most studies designed for database development are probably proprietary, and therefore not published. A notable exception is Wood et al's study, "Overlap Among the Journal Articles Selected for Coverage by BIOSIS, CAS, and Ei."¹⁰

The studies mentioned above, with the exception of Marsh's analysis (see endnote #6) and Wood's (endnote #10), have been carried out by identifying either one key journal title or a set of core journal titles for a given subject discipline. Kawasaki's analysis' of four agricultural related A&I services, (Agricola, Biological and Agricultural Index Plus, Biological Abstracts, and CAB Abstracts), found that the CAB Abstracts database had the highest coverage rate for a

set of core agricultural serials compiled in the Literature of the Agricultural Sciences (Olsen ed. 1991-1996).

This practice of identifying key or core journal titles for the evaluation of a subject is, for the most part, a reasonable simplification that can account for a majority of a subject's primary journals. However, this simplification often misses the rarely cited publications, as noted informally in the 80/20 rule, or more precisely in Bradford's Law, that states that the set of cited journal titles can be subdivided into three subsets, with one subset representing a few titles that account for a large percentage of works cited, and another subset representing many titles that are only rarely cited.¹¹ The problem of accounting for numerous journal titles that are only occasionally cited, is of particular importance to a subject discipline as diverse as agriculture.

Although the use of common knowledge or a professionally constructed bibliography are the standard methods of identify a key or set of core journal titles, one might be inclined to consider the use of the Journal Impact Factor (JIF). The Journal Impact Factor is a moderately complex form of citation analysis which is determined by citation count to a journal relative to the citation count within a given domain (typically as defined by ISI). Although particularly useful for the purposes of identifying journal influence and research patterns, the usage of JIF for collection development purposes is, at best, a derived measure of patron need. Although one can probably assume that there is a strong correlation between the needs of researchers within a given domain and the JIF for journals in that given domain, the JIF itself is *not* a measure of need. Instead of using JIF for collection development purposes, it would be best to identify a more direct measure of patron need.

Even though journal level studies can be significantly easier to conduct than article level studies, and are often well-suited for object and intent of the study, citation analysis studies that focus on format/source level cannot be as precise as studies that focus on the article level, and therefore the evaluation of A&I services should look at the finer-grained level of analysis if possible. For this study, the author decided that this level of citation analysis would be feasible, and would provide the most indicative measure of researcher need available.

Methodology

Despite the apparent representativeness of citation analysis at the article level, the library literature on the evaluation of Abstracting and Indexing services rarely employs this measure. The absence of such studies could be due to the apparent technical difficulty of methodically checking the availability of each cited article in each of databases; the process is tedious and almost impossible to automate, and yet requires a high degree of attention. However, despite these difficulties, it is believed that the methodology described in this paper will allay concerns regarding the feasibility of related studies.

The methodology section of this paper consists of two parts: first, the identification of user need, and second, the evaluation of A&I services according to this measure of user need.

Identifying User Need

In order to collect a representative sample of works cited from USDA publications, the author designed a simple sampling method that accounted for the diversity of research at USDA-ARS without artificially emphasizing one branch of research. First, the author requested all USDA ARS authored publications with publication date, submitted date, or approved date 2005, returning 3,489 publications. From this set of documents, the author then removed forty-four records with irreconcilable data extraction errors, and then dropped all records not published in 2005, leaving 2,008 records. This subset was then ordered chronologically by the date they were reported/submitted to USDA. Starting with the first record, the author selected every twentieth record resulting in 101 publications, creating a ten percent sample of USDA-ARS publications. From this sample of publications, the author extracted bibliographic citations of all works cited from either the Scopus citation or from the full-text (electronic or print), returning 3,311 works cited. the author then utilized the MSEXcel random number function to generate a 10% random sample, returning 330 works cited.

Evaluating A&I

These works cited were checked against four A&I databases, (Scopus, Agricola, Biosis, CAB), as well as for full-text availability in DigiTop, the National Agricultural Library's collection of electronic resources available to USDA. Three of the databases, (Agricola, Biosis, and CAB), were checked using WebSPIRS, Ovid Technologies cross-database search. To prevent false drops, it was decided that the following search strategy would be consistent across platforms and would cast the broadest net with the ability to narrow:

- First, the author attempted to identify the longest continuous title phrase, uninterrupted by stopwords, hyphenated words, words in quotes, abbreviations, and words that have multiple correct spellings.
- Second, the author then identified a specific title word/phrase combined with either:
 - another specific title/phrase and/or
 - truncated author last name.

If neither step retrieves the document in a database, it is assumed absent from that database. If the database returned too many results, additional words or author names, and rarely journal title, were added. Volume and issue, and as well as pagination were checked against the original citation to verify the database record against the original citation. The availability of a works cited in a database was noted as either "found" or "not found."

Findings

The author decided that the following means of analyzing the collected data would provide useful insight into the ways the databases covered the needs of USDA ARS researcher: how well a given database covered the sampled works cited, how similar a given database was to the sampled works set, coverage by format, coverage by year, uniqueness of content, and

complementary coverage. In this section, the author will describe each measure and present the findings. In the following section, the author will discuss a few interpretations.

Coverage and Similarity

The author found that an ARS researcher could have found either the citation or full-text for 297 of the 330 works cited in one of the A&I databases or full-text journals, resulting in a 90% overall coverage rate. Coverage was defined as the percentage of the works cited found in a given database. The similarity measure was designed to identify how ‘focused’ a given database is to the researcher need.

Coverage

Table1: Overall coverage of DigiTop resources

	<i>Combined</i>	<i>Agricola</i>	<i>BIOSIS</i>	<i>CAB</i>	<i>Scopus</i>	<i>Full Text</i>
Count	297	181	241	193	235	142
Percentage	90%	54.9%	73.0%	58.5%	71.2%	43.0%

Similarity

Table2: Overall similarity of DigiTop resources. Calculated by dividing ‘works cited count’ by database size.

	<i>Agricola</i>	<i>BIOSIS</i>	<i>CAB</i>	<i>Scopus</i>
Count	181	241	193	235
Database size (millions)	3.8	13	4	28
Similarity	47.6	18.5	48.3	8.4

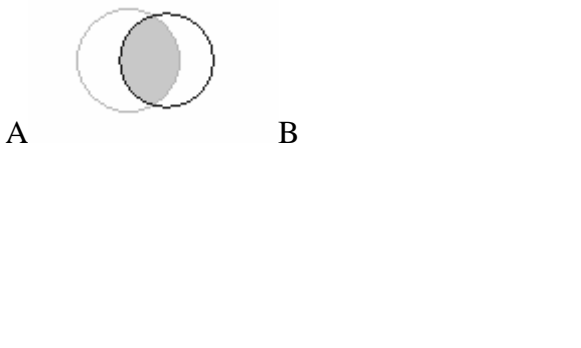
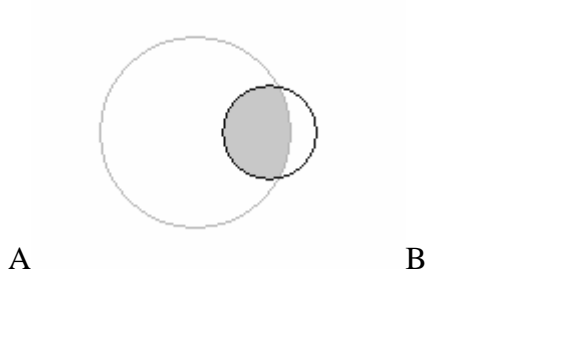
	<i>Agricola</i>	<i>BIOSIS</i>	<i>CAB</i>	<i>Scopus</i>
Measure				

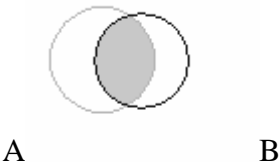
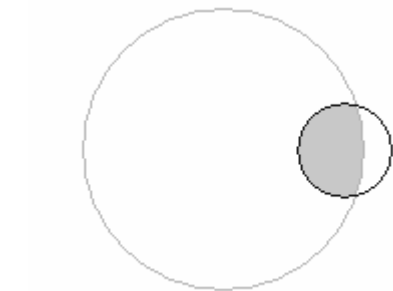
Coverage and Similarity Venn Diagrams

Coverage and similarity measures can be visually represented by Venn diagrams. For the purposes of display, database size has been scaled to 1/10,000. In the following diagrams, database coverage is represented by the shaded fraction of the circle B. Similarity is represented by the shaded fraction of circle A.

All Works Cited

Table 3: Coverage and Similarity Venn Diagrams.

<p><u>Agricola</u></p> <p>Left circle (A): Agricola 380 (Scale 1/10,000)</p> <p>Right circle (B): WorksCited sample (330 records)</p> <p>Shared area: Works Cited sample found in Agricola database (181 records)</p> 	<p><u>Biosis</u></p> <p>Left circle (A): Biosis 1300 (Scale 1/10,000)</p> <p>Right circle (B): WorksCited sample (330 records)</p> <p>Shared area: Works Cited sample found in Biosis database (241 records)</p> 
---	---

<p><u>CAB</u></p> <p>Left circle (A): CAB 400 (Scale 1/10,000) Right circle (B): WorksCited sample (330 records) Shared area: Works Cited sample found in CAB database (193 records)</p> 	<p><u>Scopus</u></p> <p>Left circle (A) : Scopus 2800 (Scale 1/10,000) Right circle (B): WorksCited sample (330 records) Shared area: Works Cited sample found in Scopus database (235 records)</p> 
--	---

Format

The format measure was designed to identify variations in how the various databases cover the various formats. The formats were defined as: journal articles, books, book chapters, proceedings, and 'Other.' Proceedings were identified by locating keywords in source title (annual, meeting, conference, proceedings, symposium, conference, etc), and then checking these source titles against cataloguing records in WorldCat (Doc Type: serial; Material type: Conference). When multiple records in WorldCat suggested different categories, the author went with the most commonly chosen format. The 'Other' category includes: computer file, database, dissertation, institutional document, map, other.

Of the 33 works cited that could not be located in one of the A&I databases or full-text journals, 82% (27) were categorized as one of the alternative (non-journal title) formats.

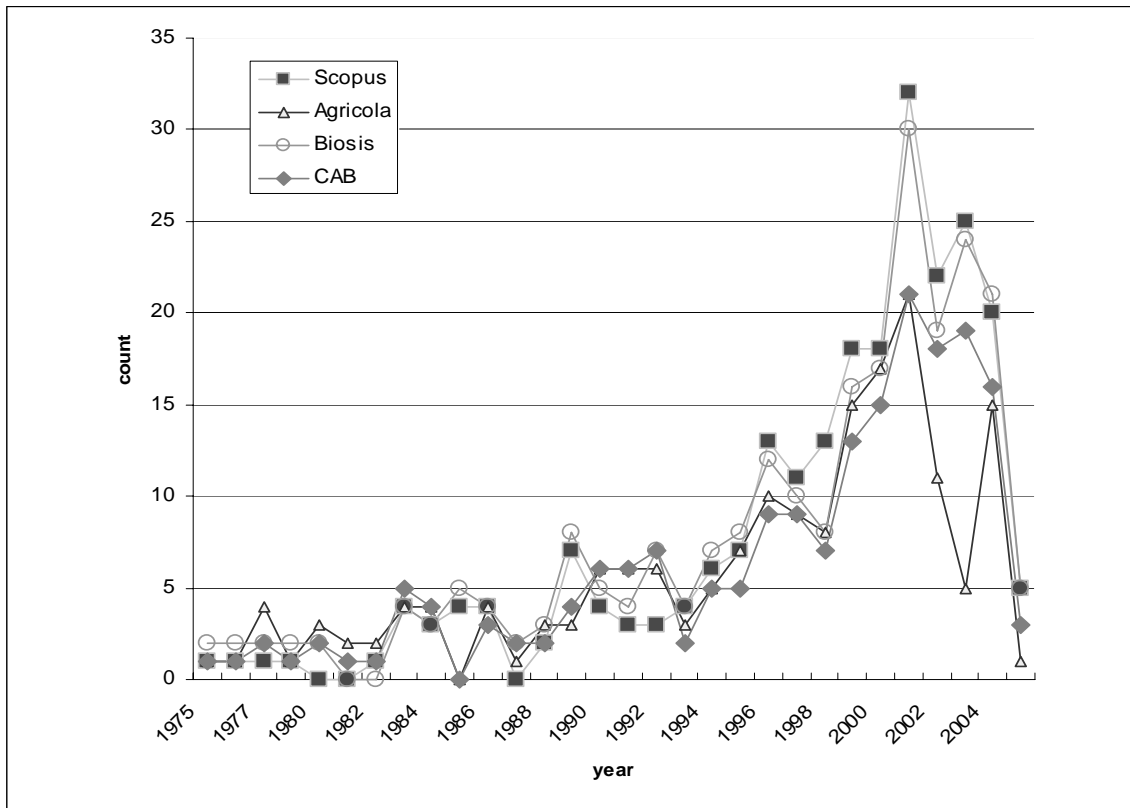
Table 4: Coverage by format.

	<i>Agricola</i>	<i>Biosis</i>	<i>CAB</i>	<i>Scopus</i>	<i>Full Text</i>
All (100%)	54.9%	73.0%	58.5%	71.2%	43%
J. Articles (83%)	57.7%	84.3%	66.4%	84.3%	51.9%
Proceedings (2.4%)	62.5%	25.0%	37.5%	12.5%	0.0%
Chapters, books (4.9%)	18.8%	37.5%	12.5%	0.0%	0.0%
Books (6.7%)	63.6%	9.1%	18.2%	13.6%	0.0%
Other (3.0%)	40%	60.0%	30.0%	0.0%	0.0%

Year

The availability of works cited in the various A&I databases (and full-text) were analyzed by year. The chart is a count of coverage by year.

Graph 1: Coverage by year. The majority of works cited were published between 2001-2003.



Uniqueness of content

Uniqueness of content was designed to identify content unique to a given database. A works cited was considered unique to the database if it was not found in the other databases. (this was done by importing the records into an MSAccess database, and then running an SQL query against this data).

Table 5: Uniqueness of content. Agricola returned the greatest number of unique records.

	<i>Agricola</i>	<i>Biosis</i>	<i>CAB</i>	<i>Scopus</i>
Records unique to database	14	7	1	13

Complementary coverage

Complementary coverage was designed to identify the similarity or dissimilarity of two databases. A works cited was counted if it appeared in both databases.

Table 6: Complementary coverage. The combination of Agricola and Scopus returned the highest percent of coverage.

	<i>Agricola and Biosis</i>	<i>Agricola and CAB</i>	<i>Agricola and Scopus</i>	<i>Biosis and CAB</i>	<i>Biosis and Scopus</i>	<i>CAB and Scopus</i>
Comple- mentary Coverage	84.2%	69.7%	86.4%	80.9%	83.0%	82.1%

Discussion

This study found that Biosis and Scopus had significantly higher rates of overall coverage than Agricola or CAB, even though the Agricola and CAB databases were found to be considerably more similar to the demonstrated patron need than Biosis or Scopus. This finding is in contrast to the findings of Kawasaki's journal-level study, which concluded that CAB provided the most agricultural content. Though not intended to be a direct comparison, with many unaccounted for exogenous variables such as differing years of study, differing intent of the study focusing on the field of agriculture and not the needs of one library, as well as slightly different set of databases, it appears as if the increased level of granularity revealed limitations to the content of the CAB A&I service, at least for ARS researcher need.

Although material published before 2000 was still highly cited, analysis by year reveals that a high percentage of works cited were published 2000 to current, with a peak around 2002-

2003. It is possible that the gap between USDA-ARS publication date and the publication dates of the material cited could be mitigated by utilizing and recognizing alternate methods of dissemination, such as open access journals and institutional repositories, which would improve the timeliness and competitiveness of USDA-ARS research.

Although Agricola's overall coverage rates were lower than either Biosis or Scopus, analysis by format revealed good coverage of formats other than journal articles, with Agricola coverage rates the highest for proceedings, chapters, books, and second highest for Other.

The findings of this descriptive study suggest that the Agricola abstracting service focus on these alternate formats, which amounted to 17% of the works cited, as well as other material unlikely to be indexed by commercial A&I services. Similarly, although the population of unique records per database is limited, this study finds that Agricola and Scopus contain more unique content than CAB or Biosis. Likewise, the A&I pair of Agricola and Scopus had the highest level of complementary coverage, followed by Agricola and Biosis. This suggests that Agricola's content complements these commercial services, and that the development of the database should continue to focus on unique content.

Conclusion

It is hoped that this study has been more than an exercise in curiosity, and that the following recommendations for the National Agricultural Library, and for the library community, will be considered.

The fact that the Scopus and Biosis abstracting & indexing databases had the highest rates of coverage and similarity is of particular note to the National Agricultural Library. This finding should be considered when making collection development, user instruction, and 'web-real estate' decisions. This study adds further evidence for the motivation behind the rescoping of the Agricola. The findings suggest that the database focus on unique content, such as alternative formats, as well as material that is unlikely to be indexed by the commercial Abstracting & Indexing databases.

In order to identify rates of improvement, problem areas, etc., it is recommended that this study be conducted on a yearly basis. Such a longitudinal study will be able to identify whether, and to what degree, the intended results of the Agricola rescope have been achieved. Specifically, we should see an overall improvement in coverage, as well as higher rates of complementary coverage and uniqueness.

For the greater library community, the author would like to reemphasize the utility of the citation analysis technique as a direct and effective method of identifying patron/researcher need. Even though this study utilized citation analysis for the evaluation of abstracting and indexing databases, this technique can be applied to the evaluation of other library services, including: journal subscriptions, monograph collections, interlibrary loans, audiovisual collections, digital libraries, and institutional repositories. Since the same dataset can be easily reused for multiple services, this evaluative technique is worth consideration.

Finally, although there are several aspects of the National Agricultural Library that set it apart from other research libraries, there is no reason why this technique cannot be applied to other research settings. For user groups at academic institutions where performance is oftentimes defined by research output, citation analysis can provide valuable insight into these groups' needs. By providing a benchmark, libraries can measure the effectiveness of changes in collection development policies, and recommend new ones.

¹ Joseph R. Kraus, "Comparing Journal Use Between Biology and Undergraduate Students," *Issues in Science & Technology* 43 (Summer 2005): 00, <http://www.istl.org/05-summer/article2.html>.

² Johan Bollen, H. van de Sompel, J.A. Smith, and R. Luce, "Toward alternative metrics of journal impact: a comparison of download and citation data," *Information Processing and Management: an International Journal*, 41, 6 (December 2005): 1419-1440.

³ Deborah D. Blecic, Joan B. Fiscella, and Stephen E. Wiberley, Jr., "Measurement of Use of Electronic Resources: Advances in Use Statistics and Innovations in Resource Functionality," *College and Research Libraries*, 68,1 (January 2007): 26-44.

⁴ Peter Jasco, "Analyzing the Journal Title Coverage of Abstracting/Indexing Databases at Variable Aggregate and Analytic Levels," *Library & Information Science Research*, 20,2 (February 2002): 133-151.

⁵ Tim Laborie, H.D. White, "Library and Information Science Abstracting and Indexing Services: Coverage, Overlap, and Context," *Library & Information Science Research*, 7,2 (April 1985): 183-196.

⁶ Spencer S. Marsh, "Bibliography of Bioethics and Index Medicus: comparison of coverage, publication delay, and ease of recall for journal articles on bioethics," *Journal of the Medical Library Association*, 75,3 (July 1987): 248-252.

⁷ Ruth E. Fenske, N.J. Mayer, "Title Coverage of Seven Indexes to Religious Periodicals, Reference & User Services Quarterly, 37,2 (December 1997): 171-197.

⁸ Jodee L. Kawasaki, "Agriculture Journal Literature Indexed in Life Sciences Databases." Issues in Science and Technology Librarianship. 40 (Summer 2004), <http://www.istl.org/04-summer/article4.html>

⁹ Marian A. Burright, T.D. Hahn, M.J. Antonisse, Understanding Information Use in a Multidisciplinary Field: A Local Citation Analysis of Neurosciences Research, College & Research Libraries, 66,3 (May 2005): 198-210.

¹⁰ James L. Wood, "Overlap Among the Journal Articles Selected for Coverage by BIOSIS, CAS, and Ei," Journal of the American Society for Information Science, 24,1 (January 1973): 25-28.

¹¹ Samuel C. Bradford, "Sources of Information On Specific Subjects," Journal of Information Science, 10,4 (1985): 173-180. Reprint of original 1934 article.