Faculty Publications   LSU Libraries

2004

Content Analysis of the Leading General Academic Databases

Kelly D. Blessinger
Louisiana State University, kblessi@lsu.edu

Maureen Olle-LaJoie
maureen.olle-lajoie@uwrf.edu

Content analysis of the leading general academic databases

Kelly Blessinger & Maureen Olle

ELSEVIER

Content analysis of the leading general academic databases

Kelly Blessinger[a],* and Maureen Olle[b]

[a]*Louisiana State University Libraries, 141 Middleton Library, Baton Rouge, LA 70803, USA*
[b]*Louisiana State University Libraries, 53 Middleton Library, Baton Rouge, LA 70803, USA*

Abstract

Title lists from the three leading general academic databases were compared to evaluate journal content. Some of the elements that were analyzed include subject coverage, the quality of titles provided, accessibility issues, and trends over time for these databases. While database title lists are now readily available on the Internet for comparison, as our article illustrates, these lists need further evaluation for proper assessment.
© 2004 Elsevier Inc. All rights reserved.

Keywords: Aggregate databases; Subject coverage; Full-text; Database comparison

1. Introduction

Academic libraries typically serve a wide variety of clientele. To suit these diverse information needs, a general academic database covering a broad array of topics is essential. Currently, the three most prominent general databases in academic institutions are EBSCO Academic Search Premier (hereafter called "EBSCO"), Gale Expanded Academic ASAP (hereafter referred to as "Gale"), and ProQuest Research Library (including the 15 subject-specific modules: Arts, Business, Children's Education, General Interest, Health, Humanities, International, Law, Military, Multicultural, Psychology, Social Sciences, Sciences, and Women's Interests) (hereafter referred to as "ProQuest"). Since these databases are similar in scope, and an overlap in coverage is assumed, most institutions will only want to subscribe

---

* Corresponding author.
E-mail addresses: kblessi@lsu.edu (K. Blessinger), molle@lsu.edu (M. Olle).

to one general database. This article will illustrate the variation in subject coverage and quality of journals provided within each of these databases. Other factors considered included full-text access to materials, unique titles that only appeared in one of the databases studied, and how these databases have grown and changed throughout their histories. This article will help institutions make informed decisions as to which database will best suit their needs.

2. Prior studies

Since full-text databases first became available as research tools, there have been numerous studies conducted to discuss issues relevant to these resources. Studies have varied from the completeness of the full text provided, to embargo periods, cancellations of serials, and comparisons of specific products. Some of these studies are discussed in the following paragraphs.

Articles that evaluated how the electronic full-text compared to the print version of journals typically focused on how complete the full-text in the databases really was. A study by Orenstein [1] reported that some features such as advertisements, letters to the editor, short articles, and graphs and charts were not available in the electronic full-text. Another study performed by librarians at the University of the Pacific Library checked 130 specific journal titles in IAC's Business Index ASAP to test for true full-text availability. The study began when they found that not all full-text that the vendor listed on their title lists was actually obtainable. The authors found that there were several cases of incorrect citations, missing issues and articles, and that over 12% of the journals listed as being full-text only had abstracts available [2].

These issues are certainly important to bear in mind when considering discontinuation of the print format. Other important issues in relation to cancellation are embargo periods and title changes/cancellations within the databases. Embargo periods are amounts of time set by the publisher where the full-text is not available electronically after the material is printed. The publishers use embargoes as safeguards so the print subscriptions are less likely to be cancelled if researchers need the most recent issues of the journals provided. Cancellations and title changes within databases are also important topics. When the access to titles can change at the whim of the aggregators and publishers, librarians have a lack of control as to what they are able to provide. The 2001 Supreme Court's ruling of *Tasini v. New York Times*, in which the court ruled that publishers must attain permission of freelance writers before including their work in electronic format affected full-text databases. While this ruling mainly affected general magazines and newspapers, any questionable content was removed from databases, thus creating holes in access for some publications. These and other issues were discussed at a roundtable discussion with the three major aggregator producers, EBSCO, Gale, and ProQuest, at the 2002 ALA conference in New Orleans [3].

A 2001 comparison study by Krumenaker [4] evaluated the three major vendors journal lists for unique titles that only appeared in one of the three databases. The study was in

response to protests from the library community due to the recent increase of periodical title holdings of EBSCOhost. It was not the large increase that upset librarians, but the idea that EBSCO might be making offers for exclusive deals with publishers to be the only vendor carrying their product. When the study was complete, it was found that the number of unique titles, meaning titles that only appeared in one of the three databases, was relatively equal, with 54% for EBSCO, 48% for ProQuest, and 42% for Gale. However, it was noted that EBSCO had a much higher rate of providing these titles exclusively, meaning the only electronic format provided, even outside of the two other databases studied. The titles that were exclusive also tended to be of a more scholarly nature. While representatives from Gale and ProQuest claimed this was due to EBSCO offering outlandish funds to secure exclusive deals, EBSCO claimed that many of these publishers were new to the electronic marketplace and that they were just the first company to ask, commenting that their contacts are not exclusive, they are just the first to provide electronic access. Another study compared the same databases in this article in regard to user end product and paid special attention to features relative to the intuitiveness of the search structure, advanced search options, and the format and availability of the full-text provided [5].

While there were some previous studies based on the evaluation of subject coverage and title evaluation across the full-text databases, they were usually category specific, such as a 2001 study that compared three full-text databases in regard to scientific coverage [6]. Another article published in 1999 focused on the social sciences content of four prominent full-text databases [7]. Since a thorough subject and content quality analysis of all of the titles within these three databases had not been published, the authors felt there was a need for this type of evaluation. Due to EBSCO's larger journal coverage and results from prior studies, the authors hypothesized that EBSCO would have the broadest coverage of the subject areas.

3. Methodology

Title lists from EBSCO, Gale, and ProQuest were downloaded from the vendors' Web sites. The lists present on each vendor's Web site in November 2002 were used for this study. The lists were reviewed, and entries with duplicate ISSNs were merged into single records. While merging entries, database holdings were noted for each title. In the case of title changes, entries were counted as unique only when the ISSN was also altered. It was assumed that if there were a title and ISSN change, there was also a significant transformation in content, such as the peer-reviewed status or a shift of focus in regard to the subject matter. Verification of each ISSN was then made in Ulrich's Periodicals Directory Online to ensure uniqueness. Titles that could not be confirmed were excluded from this study. This exclusion totaled 156 titles that could not be verified. In the final edit, there were a total of 6406 unique titles, 3980 or 62% of which were peer reviewed. This was a vast reduction from the original merged list of 10,506 entries.

The next step was to identify the peer-reviewed status and subject coverage of each title using Ulrich's Periodicals Directory. Ulrich's was chosen to provide consistency because none of the

three vendors used the same method to determine peer-reviewed status. The journal publishers provide Ulrich's with both the peer-reviewed status and subject descriptors. D. Nelson (personal communication, June 19, 2003), a representative from Ulrich's, explained that the company created its own general subject headings because the Library of Congress codes were considered too extensive and specialized for use with their directory. The first subject listed in Ulrich's was assumed to be the primary subject. The 6406 titles were divided into 375 subject groups. The subjects were then placed into the broad categories of science, social science, and arts and humanities. The assignments were determined by the subject classifications found in *Science Citation Index*, *Social Science Citation Index*, and *Arts & Humanities Citation Index*. Some subjects were not listed in any category, and likewise others straddled more than one broad category. For those subjects not listed in any category, the authors' made logical assignments based on the closest related subject. However, no determination could be made for 127 of the titles due to their broad scope. The lists provided from the citation indices contained some subjects that were divided into multiple categories (i.e., history appears in both social sciences and arts and humanities).

4. Findings

While database vendors have made their title list coverage much more accessible then they have in the past, the data provided still should be scrutinized [8]. Each of the vendor's lists included the journal titles, ISSNs, PDF image information, and dates of citation and full-text coverage. Some vendor lists provided additional information, such as the publisher information, embargo periods, and the peer-reviewed status; however, this information was not consistent throughout. The authors' combined list included a title field, ISSN field, database holdings information, peer-reviewed status, and primary and secondary subjects.

Merging the title lists of the databases was more labor intensive than initially anticipated due to inconsistencies in titles, which led to duplicate entries for single periodicals. Another complication arose with the irregularity in ISSN assignment. The database vendors listed different ISSNs for identical periodicals, sometimes listing old ISSNs for new titles. Before editing, title lists for EBSCO listed 4524 titles indexed, Gale listed 3172 indexed, and ProQuest stated that 2810 titles were indexed. After editing and verification, EBSCO provided access to 4429 titles with 3161 peer reviewed, Gale 3089 titles with 2235 peer reviewed, and ProQuest 2762 titles (2079 peer reviewed). This is a discrepancy of 95 or 2% of the titles listed for EBSCO, 83 titles or 3% of those listed for Gale, and 48 or 2% of the titles listed for ProQuest. All percentages are rounded to the nearest percent.

4.1. Subject areas

The major emphasis for all three databases in terms of subject coverage was in the social sciences with 3423 titles or 53% of the titles in this study. EBSCO had 2245 titles in the social sciences, 1494 of those titles being peer reviewed. With 1826 titles and 1133 of them

peer reviewed, Gale had slightly better coverage than ProQuest's 1708 titles (963 peer reviewed). Business and economics was the subject area with the most titles for all three databases. While Gale had superior coverage in business and economics by having 116 more titles than ProQuest and 226 more titles than EBSCO, EBSCO tended to have more titles than the other two databases in the rest of the subject areas.

Science journals had fair coverage with 2365 unique titles, which constituted 37% of the titles surveyed for this article. EBSCO's strength in the sciences was proven with 1925 titles, 1595 of which were peer reviewed, compared to Gale's 1070 titles (781 peer reviewed) and ProQuest's 827 titles (586 peer reviewed). While EBSCO dominated the majority of the science subject categories, Gale had the most titles for the subject category "sciences: comprehensive works." In subjects like biology, Gale and ProQuest had significantly lower total number of titles than EBSCO, but both databases had a high percentage of peer-reviewed titles.

Arts and humanities was the smallest category with only 885 titles, 14% of the titles in this study. EBSCO had 638 titles in the social sciences, 409 of those titles being peer reviewed. Gale had 569 titles, with 365 being peer reviewed. ProQuest had the smallest coverage with 482 titles, 267 of which were peer reviewed. Gale provided the best coverage to the largest subject area, literature. In several subjects, EBSCO had the most titles but either Gale or ProQuest had more peer-reviewed titles. Table 1 shows a further analysis of the categories, including subjects that had more than 50 titles within the categories of social science and science and more than 20 titles in arts and humanities.

4.2. JCR analysis

Citation impact factors from the Institute for Scientific Information (ISI) *Journal Citation Reports (JCR) on CD-ROM 2001 (Social Science and Science Editions)* contain citation analysis data on 7430 titles. Coverage in JCR is not comprehensive but is assumed to represent the most important journals in the social sciences and the sciences. JCR does not have an edition for the arts and humanities so journals in that category are not covered in this section of the study. Sixty percent of the titles in the Social Science JCR were contained in one of these three databases. Only 21% of the Science JCR titles were represented in these databases.

Citation impact factors track the frequency of citations. This is considered a valid measure of journal quality because frequency of citation implies scholarly acceptance [7]. The average and median impact factors were determined for the social science and science titles in each database. The average impact factor for the social science titles in JCR was 0.82. ProQuest had the highest average impact factor of the three databases with 0.96. Gale followed with an average impact factor of 0.94. EBSCO's average impact factor was 0.85. The median impact factor for the social science titles in the JCR was 0.57. Both ProQuest and Gale had a median of 0.62. EBSCO's median impact factor was 0.56. Despite having the lowest average and median impact factors, EBSCO had the highest percentage of titles in the top quartile (determined by ranking the titles by impact factor and selecting the top

Table 1
Top Subjects within the Social Sciences, Science and Arts, and Humanities

Primary subject	No. of titles	Percentage of peer reviewed (PR)	No. of titles in ProQuest	Percentage of PR in ProQuest	No. of titles in Gale	Percentage of PR in Gale	No. of titles in EBSCO	Percentage of PR in EBSCO
Social science								
Business and economics	509	51	242	50	358	54	132	50
Education	412	55	188	55	140	61	329	61
Political science	275	63	148	66	168	67	215	69
Psychology	243	89	100	98	111	95	221	89
Law	166	43	73	51	81	52	95	48
History	151	72	72	72	97	72	117	74
Sociology	128	81	63	86	80	86	113	85
Library and information sciences	81	65	43	67	23	65	66	64
Social sciences: comprehensive works	81	73	37	78	38	82	70	76
Philosophy	75	92	28	96	36	97	68	91
Linguistics	73	82	26	88	36	86	61	85
Literary and political reviews	72	42	43	37	60	38	53	45
Sports and games	71	18	44	11	24	16	25	36
General interest periodicals	70	11	54	9	31	16	30	13
Geography	52	71	16	19	27	78	40	73
Science								
Computers	246	45	84	33	115	38	163	55
Biology	229	92	56	95	56	98	205	92
Engineering	176	81	29	62	87	82	148	84
Environmental studies	98	71	23	48	49	53	77	78
Sciences: comprehensive works	78	73	33	79	58	76	48	73
Mathematics	61	95	12	92	11	100	56	95
Chemistry	56	91	4	100	13	85	50	94
Pharmacy and pharmacology	54	78	10	50	10	50	42	88
Arts and humanities								
Literature (total)	194	70	112	73	150	75	140	73
History (total)	151	72	72	72	97	72	117	74
Religions and theology (total)	134	57	54	63	62	65	115	61

Table 1 (*continued*)

Primary subject	No. of titles	Percentage of peer reviewed (PR)	No. of titles in ProQuest	Percentage of PR in ProQuest	No. of titles in Gale	Percentage of PR in Gale	No. of titles in EBSCO	Percentage of PR in EBSCO
Arts and humanities								
Literary and political reviews	72	42	43	37	60	38	53	45
Music (total)	64	48	37	38	41	59	41	61
Humanities: comprehensive works (total)	47	64	23	74	34	76	36	64
Art (total)	37	54	23	43	20	60	21	71
Archaeology	31	84	10	80	17	88	24	88
Motion pictures	28	43	20	40	17	47	17	59
Theater	26	42	16	31	18	50	19	37

25%) in the Social Science JCR at 47.7%. Gale had 47.5% of the top quartile titles while ProQuest had 45.1%. For complete information on quartiles, see Table 2.

The average impact factor for the science titles in the JCR was 1.45. ProQuest had the highest average impact factor at 2.65. Gale followed with an average impact factor of 2.56. EBSCO's average impact factor was 1.84. The median impact factor for the science titles in the JCR was 0.82. Both ProQuest and Gale had a median of 1.20. EBSCO's median impact factor was 0.96. However, EBSCO had the highest percentage of titles in the top quartile titles in the Science JCR at 21.3%. Gale had 11.2% and ProQuest had 8.5% of top quartile science titles. Seventy-five percent of the top quartile titles in the Science JCR were not in any of these databases. As shown by Table 3, EBSCO has more titles represented in both the Social Science JCR and the Science JCR. This may explain why EBSCO has lower

Table 2
Coverage of top quartile journals (by impact factor)

Database	No. of top quartile journals in Social Science JCR ($n = 421$)	Percentage of top quartile journals in Social Science JCR ($n = 421$)
Social science titles		
EBSCO	201	47.7
Gale	200	47.5
ProQuest	190	45.1

Database	No. of top quartile journals in Science JCR ($n = 1437$)	Percentage of top quartile journals in Science JCR ($n = 1437$)
Science titles		
EBSCO	306	21.3
Gale	161	11.2
ProQuest	122	8.5

Table 3
Citation impact factors

	Average impact factor	Median impact factor	No. of titles
Science citation impact factor			
Science JCR	1.45	0.82	5749
EBSCO	1.84	0.96	1065
Gale	2.56	1.20	426
ProQuest	2.65	1.20	294
Social science citation impact factor			
Social Science JCR	0.82	0.57	1681
EBSCO	0.85	0.5	795
Gale	0.94	0.62	674
ProQuest	0.96	0.62	634

average and median impact factors than Gale and ProQuest but more representation than those databases in the top quartile titles.

4.3. Full text

It is important to remember that full text does not mean full coverage. None of the three vendors listed provide cover-to-cover treatment for these titles, and that should be remembered when journal cancellation is being considered. Having that noted, EBSCO provided full text for 3602 or 80% of its titles, Gale provided full text for 1949 (61%), and ProQuest provided 1932 (69%) of its titles full text.

4.4. Unique titles

For the purpose of this study, unique titles are titles that appeared in only one of the three databases. This does not indicate that no other database outside of this study provides access to these titles. EBSCO led the field with the number of unique titles, offering 2250 unique titles, 1646 of which were peer reviewed. Gale followed, providing 910 titles with 451 peer reviewed, and ProQuest provided 743 unique titles, 199 of which were peer reviewed.

4.5. Cancellations

Each vendor's title lists were checked to ascertain consistency of access to titles. Since the titles that were totally removed will no longer be on the current title lists, these numbers do not represent those titles, just those to which some access remained. It should also be noted that the authors did not differentiate between titles that were cancelled and those titles that ceased publication. What was noted was that if a database no longer provided current access to a title, it was counted as a cancellation. The years of coverage for citation and

Table 4
Comparison of criteria

Criteria	EBSCO %	Gale	ProQuest %
Total indexed	4429	3089	2762
Full-text [a](percentage of total provided in database)	3602 (80%)	1949 (61%)	1932 (69%)
Peer reviewed (percentage of total provided in database)	3161 (71%)	2235 (72%)	2079 (75%)
Unique titles	2250	910	743
Unique peer-reviewed titles (percentage of unique titles)	1646 (73%)	451 (50%)	199 (27%)
Arts and humanities titles (percentage of total in category, $n = 885$)	638 (72%)	569 (64%)	482 (55%)
Science titles (percentage of total in category, $n = 2365$)	1925 (81%)	1070 (45%)	827 (35%)
Social science titles (percentage of total in category, $n = 3423$)	2245 (66%)	1826 (53%)	1708 (50%)
Start of citation coverage [a]	1790	1980	1970
Start of full-text coverage [a]	1904	1980	1966
Citation cancellations [a] (percentage of total provided in database)	497 (11%)	402 (13%)	383 (14%)
Full-text cancellations [a] (percentage of total provided in database)	359 (8%)	430 (14%)	416 (15%)

Note. Figures all rounded to the nearest number.
[a] These figures were tabulated using the vendor's original lists and not the edited list.

full-text cancellations are different, and they were treated as different entities when determining the number of cancellations. Of the full-text cancellations, EBSCO listed 359 cancellations or 8% of their total offerings; Gale listed 430 cancellations or 14% percent; and ProQuest listed 416 cancellations or 15% of their titles. For the citation cancellations, EBSCO listed 497 titles cancelled (11%); Gale listed 402 titles (13%); and ProQuest listed 383 titles, which constituted 14% of their offerings (Table 4 illustrates this and other important findings).

5. Trends over time

5.1. Years of coverage

Since the years of coverage for citation and full text are different, they were treated as different entities when determining the years of coverage. For citations, EBSCO is the only database that had any indexing/abstracting for the years 1790–1969, for which it provides this service for 53 titles. ProQuest starts providing citation information in 1970, while Gale does not provide any coverage until 1980. Gale took the lead over EBSCO for indexing starting in 1981–1990, but then EBSCO eclipsed both Gale and ProQuest's citation numbers from 1991 to 2002, with more than double the number of journal title indexing from 1996 to 2000.

For full-text coverage, EBSCO is the only database that provides access in full-text format for journals before 1980. EBSCO provides full-text access to 175 titles between 1904 and 1980, while Gale and ProQuest only have one title each, respectively. Gale and ProQuest took a slight lead over EBSCO in coverage with full-text starting in 1991–1995, but EBSCO again offers the most full-text journals consistently from 1996 to 2002, with major journal title additions from 1996 to 2000. All told, EBSCO provides 46% more full
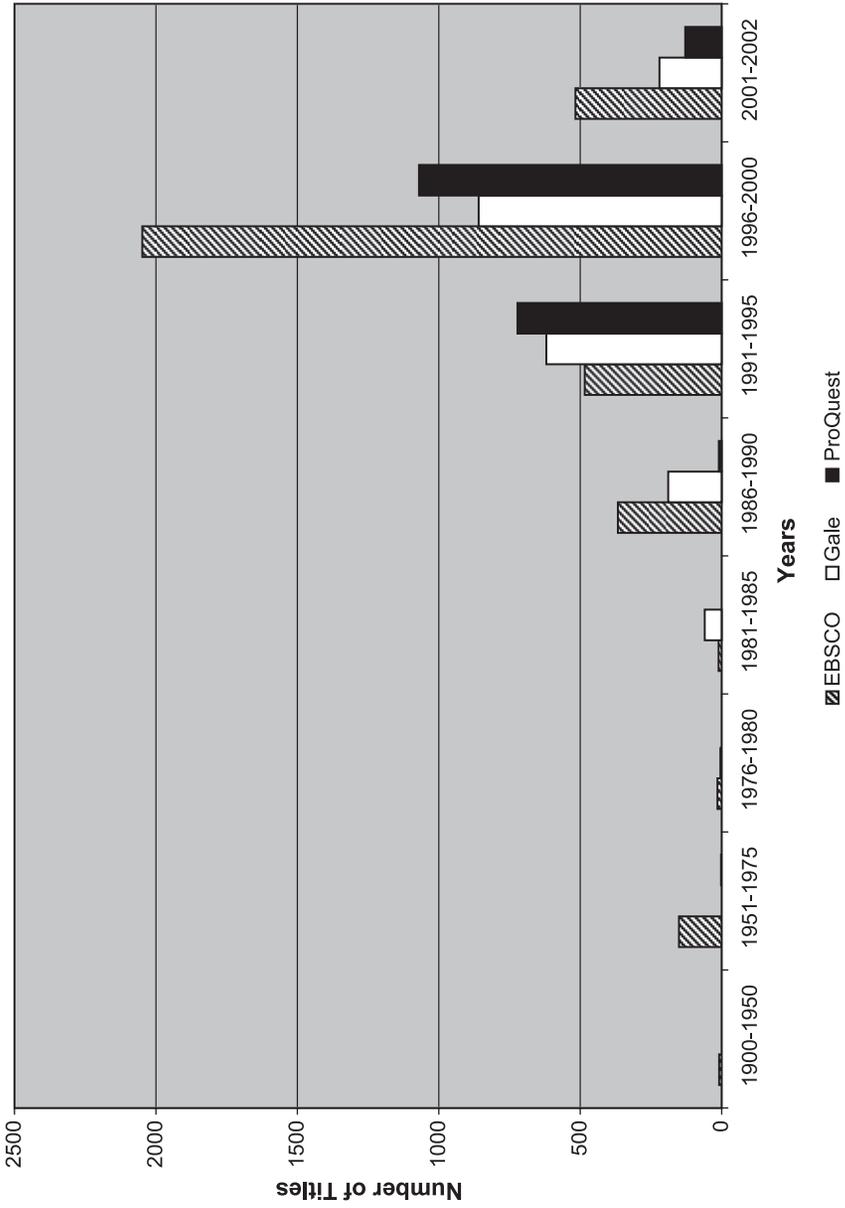
Fig. 1. Start of full-text coverage.

text than Gale, and 47% more full-text than ProQuest. See Fig. 1 for a graphical display of this information.

5.2. Recent trends

Since the dawn of full-text databases in the late 1980s, there has been tremendous growth. According to the 1987 article by Pagell [9], the amount of journals available in full text at that time numbered in the hundreds. Since the study was started, each database has made changes to their title lists. In November 2003, the authors' acquired new title lists for each database and compared these to the lists from November 2002. EBSCO added 3213 titles during that year, which represents an increase of 73%. Of these titles, 2993 entries did not appear in either Gale or ProQuest's new lists. Gale added 177 unique titles and 241 total titles for an 8% increase. Finally, ProQuest added 199 new titles, a 7% increase, with 123 of those being unique titles.

Each database had a number of deletions from their titles lists as well. EBSCO removed a total of 36 titles, 14 of which neither Gale nor ProQuest had. Gale removed 5 titles, 3 of those titles being unique to that database. ProQuest deleted 26 titles from their list, 10 of which were not available in EBSCO or Gale. Despite the overlap in databases, there were no single title deletions that were done in more than one database. It was noted that EBSCO removed quite a few titles in the business and economics field.

As a follow-up, the authors contacted each of the vendors to determine pricing differences and future title acquisition plans. None of the vendors were willing to disclose pricing information. Some vendors stated that it was company policy to not publish database costs due to the high number of factors that go into determining the price.

Since 1996, Gale has been adding mainly peer-reviewed titles to Expanded Academic [10]. Non-peer-reviewed titles are added only when the full text becomes available for a publication that they had previously indexed. Since January 2004, 15 titles were added to Gale. While the company focuses on publishers rather than subjects when making additions to this database, the new titles for 2004 will cover literature and science.

In January 2004, EBSCO added 606 titles [11]. Academic Search Premier will be displaying title changes for journals and allowing users to search on the subject or description of a journal. No plans for future title acquisitions were mentioned.

ProQuest solicits additions by focusing on a target list of 400 titles from society-based publishers. The target list concentrates on scholarly journals to which there is no electronic access. This year, ProQuest will be adding titles from education and the social sciences [12].

6. Conclusion

The reader is reminded that this study is a snapshot in time of these databases, using data that were collected in November 2002. Each of the databases analyzed have dynamic compositions, as evidenced by the number of titles added and deleted by each database

between 2002 and 2004. Any reader is urged to do current comparisons before making a purchase decision. There are important factors to consider when determining database selection. Main areas to contemplate include subject coverage, quality of titles, and years of coverage. Proving our hypothesis to be accurate, EBSCO provided the most in-depth access with the largest number of titles, the greatest years of coverage, and the most peer-reviewed journals of the three databases. They have also been the most aggressive in adding new titles in recent years. Knowing the current status of each of the databases, it might be useful for institutions to use our study as a base and evaluate the changes to journal coverage on an annual basis to determine the growth or stagnancy of the individual products.

References

[1] Orenstein, R. (1993). "How full is full" revisited: A status report on searching full-text periodicals. *Database Magazine*, *16*(5), 14–22.

[2] Grzeszkiewicz, A., & Hawbaker, A. C. (1996). Investigating a full-text journal database: A case of detection. *Database*, *19*, 59–62.

[3] Albanese, A. R. (2002). An LJ round table with the aggregators. *Library Journal*, *127*(5), 34–38.

[4] Krumenaker, L. (2001). A tempest in a librarian's teapot: EBSCO, ProQuest, Gale Exclusive, and unique titles. *Searcher*, *9*(7), 40–45.

[5] Blessinger, K., & Olle, M. (2003). Comparison of three primary aggregator databases. *Serials Librarian*, *45*(1), 53–58.

[6] Hill, J. B. (2001). Aggregated science: An examination of three multi-disciplinary databases. *Issues in Science and Technology Librarianship, vol. 30*. Available: http://www.istl.org/01-spring/article4.html

[7] Black, S. (1999). An assessment of social sciences coverage by four prominent full-text online aggregated journal packages. *Library Collections, Acquisitions & Technical Services*, *23*(4), 411–419.

[8] Still, J. M., & Kassabian, V. (1999). Selecting full-text undergraduate periodicals databases. *EContent*, *22*(6), 57–62.

[9] Pagell, R. (1987). Searching full-text periodicals: How full is full? *Database*, *10*, 33–36.

[10] Hawkwood, B. Personal communication with the authors, May 13, 2004.

[11] Petrakis, C. Personal communication with the authors, May 14, 2004.

[12] Hasan, M. Personal communication with the authors, May 20, 2004.