Automated Collection Analysis Using the OCLC and RLG Bibliographic Databases

Nancy P. Sanders, Edward T. O'Neill, and Stuart L. Weibel

This study examined the feasibility of automating the labor-intensive process of collection analysis. Collections in botany and mathematical analysis from institutions holding membership in the Committee on Institutional Cooperation (the Big Ten universities plus the University of Chicago) served as the study population. The databases of the Online Computer Library Center (OCLC) and the Research Libraries Group (RLG) were initially used as the sources of holdings information. The study found that the methodology provided a promising alternative means of analyzing and comparing library collections. However, due to varied cataloging practices of the participating libraries, accurate results could not be obtained without local verification of the holdings data.

The growing trend toward research library participation in cooperative collection development agreements has prompted collection managers to seek consistent means to evaluate and compare their collections. Unfortunately, most methods currently available are labor-intensive. The purpose of this study was to test the feasibility of using the databases of the bibliographic networks for computerized collection analysis to reduce the labor required.

The project was formally initiated in summer 1985 when the Online Computer Library Center (OCLC) and the Research Libraries Group (RLG) were invited by the Committee on Institutional Cooperation (CIC) to participate in a meeting of its science bibliographers and collection development officers. The meeting explored the potential for cooperative collection development among CIC institutions. Discussions with participants and program planners suggested that science and technology collections would be good subject areas for study.

The CIC meeting was held at the University of Chicago, September 12 and 13, 1985. During that meeting, some preliminary analyses for the OCLC member libraries were presented. Following the discussion, it was decided to expand the study to include all CIC member universities: Chicago, Illinois, Iowa, Michigan, Minnesota, Wisconsin, Indiana, Michigan State, Ohio State, and Purdue.

Nancy P. Sanders is Head, Home Economics Libraries, at Ohio State University Libraries, Columbus, Ohio 43210-1285. Edward T. O'Neill is Senior Research Scientist, and Stuart L. Weibel is Associate Research Scientist at OCLC Online Computer Library Center, Dublin, Ohio 43017-0702. *The authors thank the Research Libraries Group, particularly Leslie Hume, for assistance and support. We also thank all CIC libraries for their help. It would have been impossible to complete this study without the detailed local verification provided by the staffs of the individual libraries.*

A literature review revealed substantial work in the area of collection analysis, particularly in collection overlap. This previous work is summarized in William Gray Potter's review of relevant research.[1] Much of the work completed in the 1960s and 1970s investigated the feasibility of establishing processing centers, union catalogs, or cooperative collection development agreements.

Several overlap studies based on methodologies different from that planned for this study were examined for relevant findings. Many earlier studies were based on random sampling from card catalogs or shelflists. For example, William Nugent's study of six New England state universities;[2] Ellen Altman's investigation of the optimum composition of a secondary school interlibrary loan system;[3] William Cooper, Donald Thompson and Kenneth Weeks' study of overlap in the University of California system;[4] and Edward O'Neill and Mary Lynn Seanor's analysis of the library collections in western New York State[5] take this approach.

Later studies such as those by Thomas Nisonger of the libraries in north Texas;[6] Barbara Moore, Tamara Miller and Dan Tolliver at the University of Wisconsin;[7] and Glyn Evans, Roger Gifford, and Donald Franz in New York State[8] employed OCLC archive tapes in collection analysis. Potter used the LCS library computer network in Illinois academic institutions.[9] While these studies, based on comparisons of random samples rather than recommended lists, were of interest, the methodologies and populations were sufficiently dissimilar to render comparisons difficult. The potential problems common to overlap studies in general were well documented by Michael Buckland, Anthony Hindle, and Gregory Walker.[10]

SAMPLING METHODOLOGY

Random samples of 500 monographic records from each of the two subject areas, botany and mathematical analysis (which includes calculus, functional analysis, functions, and differential equations), were extracted from the OCLC Online Union Catalog. These two subject areas were selected because their bibliographic character provides a useful contrast, each was collected by all of the institutions, and they were readily identifiable subjects in both the Library of Congress and the Dewey Decimal classifications. The samples were intended to be representative of recently published monographs in the subject areas, and thus in the pool of potential library acquisitions. As such, they should not be viewed as a checklist of desirable books.

Only records with a copyright or publication date between 1978 and 1983 were included. This eliminated differential rates of retrospective conversion among the libraries as a factor in the comparison of holdings and minimized the effects of delayed acquisitions or cataloging backlogs. A book was categorized as mathematical analysis if it had a Library of Congress classification number in the range QA300–433. Books without a Library of Congress class were included if they had been classified as 515 in the Dewey Decimal classification. For botany, sample selection was based on the Library of Congress classification QK and the corresponding Dewey Decimal classification 581. At the time the samples were extracted, the OCLC Online Union Catalog contained 2,301 mathematical analysis titles and 5,044 botany titles published during the six-year period included in the study.

The sample records were then compared to related records in the OCLC Online Union Catalog to determine whether any represented a publication with substantially the same content or "text" as defined by Patrick Wilson to differentiate between the content of a work and its physical form.[11] For example, under this definition, a dissertation in photocopy, microform or typescript is considered to be a single text.

If records for a duplicate text were located, an experienced searcher determined whether the sample record was the first added to the Online Union Catalog as determined by its position in the OCLC number sequence. Only the lowest numbered record for any text was included in the sample. If the sample record was the first in the Online Union Catalog and others were added later, all library holdings symbols attached to the subsequently added records were added to the holdings of the original

sample record. This procedure maintained the statistical validity of the sample, ensuring that each text had an equal chance of being included in the sample regardless of the number of records in the database representing that text.

Different editions were considered different texts with the exception of "editions" from Latin America and non-English-speaking Europe where so-called editions are most often "printings." Therefore, different "editions" from these countries were considered to be the same text, and the records were collapsed or eliminated based on their OCLC number unless there was evidence of revision. Translations were considered to be distinct texts.

Obvious serial (not monographic series) articles that had been cataloged separately and entered as monographs were also eliminated from the sample. In most cases, determining whether two records represented the same item was not simple. Some decisions were later found to be erroneous when bibliographers examined their local records or an item in hand. These errors simply point out the problem long recognized by those who catalog in an online environment: determining whether an existing record represents a work in hand is often difficult, if not impossible, given the idiosyncrasies and lack of standardization in the publishing industry and the impossibility of adequately describing an item to distinguish it from different, though similar, works using current cataloging criteria. Following the manual search of the database and the elimination of records not representing unique texts, 392 records remained in the botany sample and 454 in the mathematical analysis sample.

As the sample was searched, all relevant OCLC holdings data were appended to the selected bibliographic records. However, because only six of the eleven CIC institutions (Illinois, Indiana, Michigan State, Ohio State, Purdue, and Wisconsin) are OCLC members, not all CIC member holdings were represented. Four of the institutions (Iowa, Michigan, Minnesota, and Northwestern) are RLG members. Chicago is not associated with either bibliographic network. To obtain holdings data for the RLG members, a listing of the bibliographic records in the samples was sent to RLG where it was checked against that database.

To test the completeness of the networks' holdings data, the OCLC holdings information was compared with local records at Ohio State. It was obvious that the networks' holdings data were incomplete, largely due to local cataloging practice. Frequently, Ohio State had cataloged an item by attaching its holdings symbol to a series or serial record, rather than adding it to the record for the individual item or "subunit." This finding highlights the problems that arise when databases designed for one purpose, in this case cataloging, are used for a different purpose, such as collection analysis.

To determine the magnitude of the "subunit" problem, the searcher located the record for each serial or series that was cited within the subunit monographic record. Each was examined to determine whether it would be feasible to assume that a given library held the monographic item if the library's symbol was attached to the serial holding record (assuming that the library holding symbol was not attached to the monographic record). In the majority of cases it was decided that it was not possible to assume this because many series contained several hundred to more than one thousand associated monographs.

As a result of this early finding and the number of records belonging to this subunit category, the libraries were asked to verify their holdings. At the same time, Chicago was asked to identify the materials held. All of the institutions agreed to check the sample against their catalogs. However, due to local difficulties, the botany sample could not be verified at Michigan and Northwestern. The Michigan data provided by RLG were used without validation, recognizing that the botany holdings for Michgan are underestimated. Because Northwestern had only recently begun entering records into the RLG database, its unverified holdings were known to be seriously underrepresented. Therefore, its botany data were excluded from the analysis.

The results of the local verification, shown in figures 1 and 2, confirm the earlier suspicions that holdings indicated by

"Holdings indicated by the bibliographic networks may not accurately reflect a library's collection; therefore, the records for the bibliographic networks should be only one of several sources used to measure collection strengths."

the bibliographic networks may not accurately reflect a library's collection; therefore, the records for the bibliographic networks should be only one of several sources used to measure collection strengths. The reason for the discrepancies vary. For example, the holdings discrepancy figures show all of Chicago's holdings as "added by the library" because records from bibliographic networks were not available at the time the sample was taken. Also, data for Minnesota underrepresented their botany holdings because the wrong holdings symbol was used during data extraction. Local cataloging practices may account for other variations, such as the "subunit" problem noted above, but further examination and explanation await future research.

HOLDING PATTERNS

Of the analyses developed from the various holdings data, three focus on individual libraries' holdings. Five examine collections of the CIC institutions as a whole and provide an overview of the potential for cooperative collection development in the consortium. Of particular interest is the frequency distribution for the number of libraries per title that shows the duplication patterns for both botany and mathematical analysis (figure 3). The distributions reflect both differences in the character of the samples and collecting patterns in the two disciplines: 44% of the botany titles and 22% of the mathematical analysis titles are not held by any of the CIC institutions. An examination of the sample titles and associated holdings patterns suggests a reason for the higher botany figure: the botany sample contains a significant number of publications on regional flora and fauna that are collected primarily by libraries in the geographic area covered.

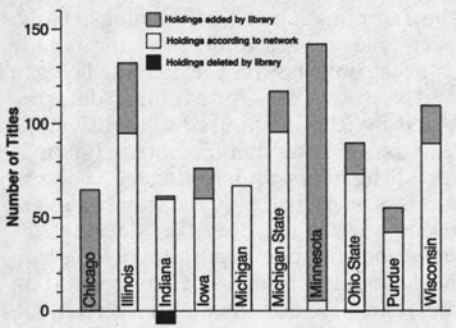

FIGURE 1

Botany Holdings

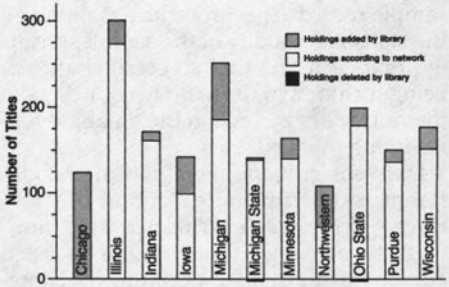

FIGURE 2

Mathematical Analysis Holdings

From the analysis of items held by only one institution, it is clear that the collections in mathematical analysis and botany lack uniqueness. Contributing factors may be the classifications examined, i.e., the subject areas may not lend themselves to specialization, or more likely, selection of a narrower classification range would be required to identify unique collections. Similar problems have been voiced by those using the classification ranges specified for the RLG Conspectus, adopted by the Association of Research Libraries (ARL) for the North American Collections Inventory Project (NCIP). They have been criticized as too broad to describe adequately the strengths and weaknesses of major research collections.

For titles held by 2 or more libraries, the percentages are similar for botany and mathematical analysis and the distribution

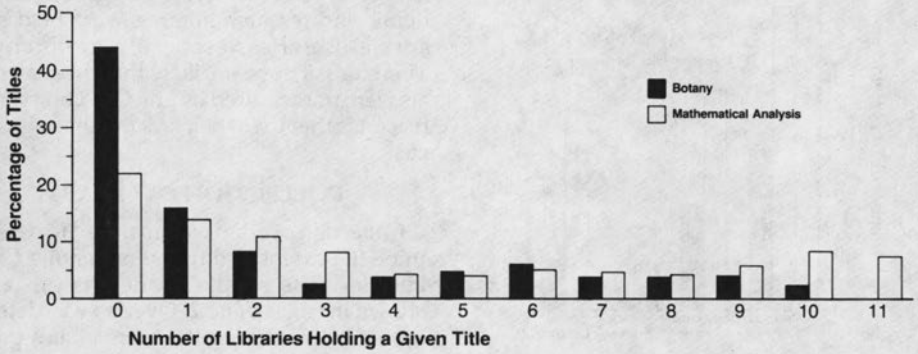

FIGURE 3

Title Duplication Patterns

is relatively flat. The number of items held by multiple libraries reflects the similarity among the collections. In botany, 25% of the sample was held by 5 or more of the 10 participating libraries. For mathematical analysis, 41% of the sample was held in 5 or more collections, and 36% was held in 6 or more libraries, indicating greater similarity among the mathematical analysis collections. The average number of libraries holding each title also indicates a greater duplication of the mathematical analysis material. Mathematical analysis items were held by an average of 4.2 CIC libraries, while the botany items were held by an average of only 2.3 libraries. Even when Northwestern's mathematical analysis holdings are excluded—to be consistent with botany—the average mathematical analysis book was still held by 4 libraries.

The pool of available materials was quite different for mathematical analysis and botany. During the period of study, approximately 350 books were published annually in mathematical analysis and 660 in botany. However, a greater proportion of the mathematical analysis materials was acquired. The CIC libraries each acquired an average of 134 books annually in mathematical analysis and 152 books in botany. The higher acquisition rate from a relatively small pool of available materials could potentially explain the higher duplication rate for mathematical analysis.

An analysis of titles not held by any CIC institution was undertaken as a result of numerous comments from CIC participants that the sample was not representative of research collections because it included many popular books, texts, and other nonresearch materials more suitable for public or school libraries. While the sample had been intended as a selection of all material published in the subject areas, the investigators questioned whether the material not held by the CIC institutions would be generally considered to be "research material." To address that question, the types of libraries holding the sample items not held by a CIC institution were analyzed. The findings are shown in figures 4 and 5.

For this analysis, a research library was defined as a member of the ARL, and academic libraries were defined as all other college and university libraries. The public libraries group also includes processing centers, school libraries, and state libraries. Only North American library holdings were included in the analysis. The examination showed that 61% of the 101 mathematical analysis titles and 60% of the 176 botany titles not held by CIC institutions were held by a least one other research library. Also notable is the number of items not held by a CIC institution that were held only by another research library: 45% in the mathematical analysis sample and 32% in the botany sample. In all cases, the sample items were more often held by research libraries than by any other type of library.

Other academic libraries held the second

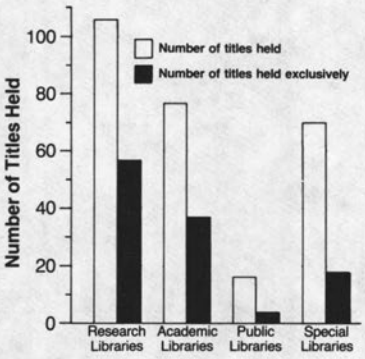

FIGURE 4

Libraries Holding Botany Titles not Held by CIC Institutions

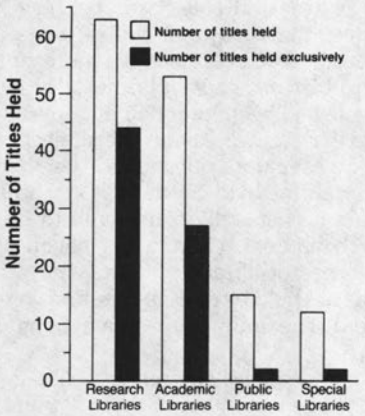

FIGURE 5

Libraries Holding Mathematical Analysis Titles not Held by CIC Institutions

largest portion of the titles not held by a CIC institution, followed by public and special libraries. For mathematical analysis, academic libraries held 52%, exclusively 28%; public libraries held 17%, exclusively 2%; and special libraries held 10%, exclusively 2%. For botany, academic libraries held 44% of the titles, 21% of them exclusively; public libraries held 10% of the titles, 2% of them exclusively; and special libraries held 36%, exclusively 10%. Thus, almost all of the materials not held by academic and research libraries were held by special libraries, especially in botany. Therefore, it appears likely that the materials were not acquired by any CIC library for reasons other than their lack of scholarly focus.

COLLECTION OVERLAP

Collection overlap was analyzed to determine the extent of duplication among CIC libraries. The results of the analysis are shown in tables 1 and 2. Overlap was determined by measuring the number and proportion of titles held in common by pairs of CIC libraries, i.e., by each CIC library compared sequentially with every other CIC library. The number held in common is shown in tables 1 and 2 below the diagonal space while the percentage appears above.

Percentages were calculated by first determining the number of volumes in the sample that were held by paired institutions (e.g., 89 + 109 or 198, in the case of the Ohio State and Wisconsin botany collections). The number of duplicated items was then subtracted (198 − 67 = 131 in the example), leaving the number of titles held by the two libraries. The number of titles held in common was divided by the number of titles held, yielding the percentage of titles held in common by the two libraries (67 / 131 = 0.511, or 51.1%).

A related research project by Charles Davis and Deborah Shaw[12] suggests that overlap is predictable by collection size. In the present study a significant positive correlation ($r = 0.58$) was found between overlap and number of volumes held by both institutions for botany. In mathematical analysis, however, there was no significant correlation ($r = -.01$). The botany finding does not support the Davis and Shaw study. However, the method of computing the overlap was different and could account for the inconsistency. Further research is required to understand the relationship between collection size and overlap.

The overlap percentages were, on average, higher for mathematical analysis than for botany. The differences are likely due, at least in part, to factors noted earlier: the

TABLE 1
COMMON HOLDINGS IN BOTANY

Institution	No. of Titles Held	No. of Common Titles/% of Common Titles									
		Chicago	Illinois	Indiana	Iowa	Michigan	Michigan State	Minnesota	Ohio State	Purdue	Wisconsin
Chicago	65	—	37.8	37.0	39.6	33.7	35.8	34.4	35.1	31.9	42.6
Illinois	132	54	—	38.8	47.5	36.6	52.8	47.3	51.4	32.6	60.7
Indiana	61	34	54	—	42.7	39.6	42.4	32.7	45.6	36.5	46.6
Iowa	76	40	67	41	—	42.0	53.2	43.4	48.6	29.7	50.4
Michigan	66	33	53	36	42	—	43.0	38.7	44.9	30.1	36.7
Michigan State	117	48	86	53	67	55	—	52.4	54.9	31.3	60.3
Minnesota	142	53	88	50	66	58	89	—	45.3	31.3	47.6
Ohio State	89	40	75	47	54	48	73	72	—	37.1	51.1
Purdue	55	29	46	31	30	28	41	47	39	—	36.7
Wisconsin	109	52	91	54	62	47	85	81	67	44	—

TABLE 2
COMMON HOLDINGS IN MATHEMATICAL ANALYSIS

Institution	No. of Titles Held	No. of Common Titles/% of Common Titles										
		Chicago	Illinois	Indiana	Iowa	Michigan	Michigan State	Minnesota	North-western	Ohio State	Purdue	Wisconsin
Chicago	124	—	37.5	58.1	52.9	44.1	51.4	55.1	46.8	47.0	52.2	50.5
Illinois	301	116	—	47.6	42.4	70.7	44.0	46.4	32.7	58.9	42.7	48.4
Indiana	170	108	152	—	57.6	55.7	55.5	59.3	45.5	51.4	56.2	57.1
Iowa	142	92	132	114	—	49.8	54.6	57.2	48.8	50.4	59.7	56.4
Michigan	252	115	229	151	131	—	48.9	54.9	38.4	59.0	51.1	56.6
Michigan State	141	90	135	111	100	129	—	57.5	46.4	52.7	54.0	58.3
Minnesota	163	102	147	124	111	147	111	—	48.9	53.0	58.2	61.2
Northwestern	105	73	100	86	81	99	78	88	—	40.3	50.9	45.3
Ohio State	198	103	185	125	114	167	117	125	87	—	51.3	49.4
Purdue	147	93	134	114	108	135	101	114	85	117	—	56.6
Wisconsin	174	100	155	125	114	154	116	128	87	123	116	—

smaller body of mathematical analysis material published and the geographical specificity of some botany material. It is highly likely that institutional collection policies also affected the overlap patterns, though this was not explicitly examined in the study.

COMPOSITION OF THE COLLECTIONS

Language of publication was found to be a useful attribute for characterizing the literature of a given subject field and for distinguishing the collecting policies of research libraries. Figure 6 shows the proportion of foreign-language material held by each institution. As might be expected, the majority of each library's collection was in English. The larger collections contain a higher proportion of non-English material. This generalization proves stronger for the mathematical analysis sample, in which the total collection size and the proportion of the foreign-language collection are closely correlated. In the botany sample, Chicago, Michigan, and Ohio State have higher percentages of non-English material than would be predicted by their comparative collection sizes. Michigan's figure is probably explained by under representation of its collection; Chicago's by its heavy research emphasis; and Ohio State's by its Herbarium staff's interest in Latin America and resulting purchases in Spanish and Portuguese, and the emphasis on Systematics for which the Biological Sciences Library purchases in many foreign languages.

The foreign-language composition of the sample, shown in figure 7, provides yet another means of illustrating the differences in the character of the two samples. The non-English portions of the mathematical analysis sample were primarily German

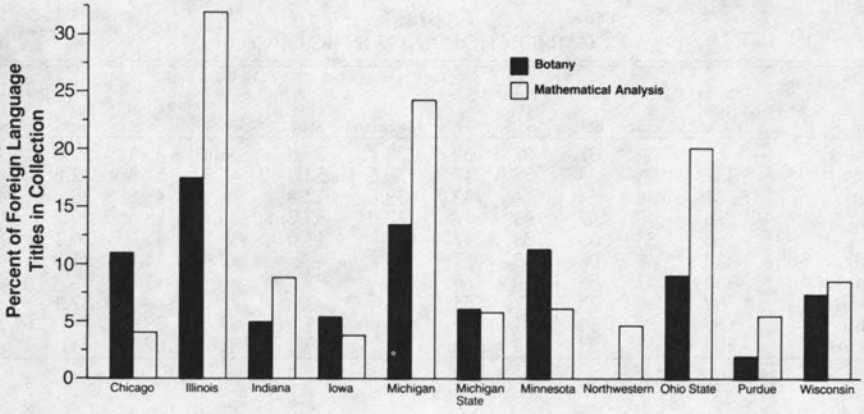

FIGURE 6

Foreign Language Holdings

(46%), Russian (30%) and French (16%). The botany foreign-language material was published more frequently in French (35%), German (24%), Russian (13%), and Spanish (10%).

IMPLICATIONS

The concept of analyzing library collections by comparing their current acquisition patterns to the pool of available monographs was found to be a viable approach to collection evaluation. Although the resulting data could be used either to compare the relative strengths of different subject areas within a single library or to compare relative strengths in a given subject among different libraries, the investigators believe that the absolute numbers are far less significant than the relative numbers. Knowing that a library acquires 25% of all available material tells little about the strength of the collection. It is only when the acquisition rate is compared to that of peer institutions that the assessments become meaningful. For example, by comparing all CIC collections, it became clear that a library acquiring 30% of the available botany material is likely building a strong collection. However, in mathematical analysis, the acquisition of 30% of the available material would produce only an average collection. Further research would be required to build a basis of comparative data for other subject areas.

"The concept of analyzing library collections by comparing their current acquisition patterns to the pool of available monographs was found to be a viable approach to collection evaluation."

The acquisition patterns for both botany and mathematical analysis materials indicate a considerable potential for coopera-

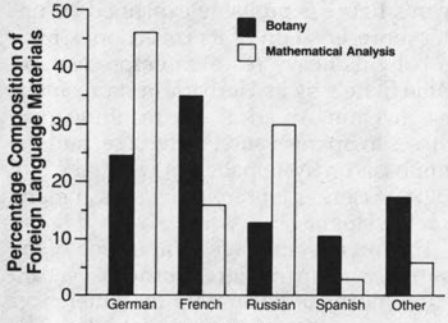

FIGURE 7

Common Languages

tive collection development among the CIC institutions. Since only approximately 5% of the acquisitions are unique, a relatively small shift in acquisition patterns could result in a significant reduction in the amount of material not acquired by any CIC institution. The result of such changes in collection development policies would be that library users would experience a small decrease in the proportion of their needs met locally, but a higher proportion would be met within the consortium. Whether the overall results of such changes would be desirable would depend on usage patterns, local expectations, and political conditions, none of which was examined in this study.

The relation between collection size and overlap bears further investigation. If such analysis could substantiate a strong positive correlation between size and overlap, then libraries contemplating cooperative agreements might rely with some confidence on the more easily obtainable collection size statistics for a particular subject classification rather than computing common holdings. Of equal importance in such a study would be a careful analysis of the collections that do not conform to the model to derive an explanation of their uniqueness.

From the analysis of the holdings, it is clear that local library cataloging practices and bibliographic networks' policies affect the utility of the online databases for collection analysis. The responses from the CIC institutions indicate a pattern of cataloging practices that require local validation to achieve reliability. Cataloging policies that resulted in partial cataloging of monographic series and no cataloging for some reserve, technical report, and theses collections became apparent in this study.

Potential uses for the results of comparative collection data include accreditation reports, collection strength analysis for proposed new programs, cooperative project viability, and Conspectus or NCIP work sheet validation. However, unless a method can be found to compensate for the unreported holdings, local validation of the holdings data is necessary to obtain consistent and reliable results. The expense of that process obviously limits its application to selected subject areas.

REFERENCES

1. William Gray Potter, "Studies of Collection Overlap: A Literature Review," *Library Research* 4:3–21 (Spring 1982).
2. William R. Nugent, "Statistics of Collection Overlap at the Libraries of the Six New England State Universities," *Library Resources & Technical Services*. 12:31–36 (Winter 1968).
3. Ellen Altman, "Implications of Title Diversity and Collection Overlap for Interlibrary Loan among Secondary Schools," *Library Quarterly* 42:177–94 (Apr. 1972).
4. William S. Cooper, Donald D. Thompson, and Kenneth R. Weeks, "The Duplication of Monograph Holdings in the University of California Library System," *Library Quarterly* 45:253–74 (July 1975).
5. Edward T. O'Neill and Mary Lynn Seanor, *A Survey of Library Resources in Western New York* (Buffalo, N.Y.: Western New York Library Resources Council, 1971).
6. Thomas E. Nisonger, "Editing the RLG Conspectus to Analyze the OCLC Archival Tapes of Seventeen Texas Libraries," *Library Resources & Technical Services* 29:309–27 (Oct./Dec. 1985).
7. Barbara Moore, Tamara J. Miller, and Don L. Tolliver, "Title Overlap: A Study of Duplication in the University of Wisconsin System Libraries," *College & Research Libraries* 43:14–21 (Jan. 1982).
8. Glyn T. Evans, Roger Gifford, and Donald R. Franz, *Collection Development Using OCLC Archival Tapes* (Washington, D. C.: Office of Education, Office of Libraries and Learning Resources, ED 152 299, 1977).
9. William Gray Potter, "Collection Overlap in the LCS Network in Illinois," *Library Quarterly* 56:119–41 (Apr. 1986).

10. Michael K. Buckland, Anthony Hindle, and Gregory P. M. Walker, "Methodological Problems in Assessing the Overlap between Bibliographical Files and Library Holdings," *Information Processing and Management* 11:89–105 (Aug. 1975).
11. Patrick Wilson, *Two Kinds of Power: An Essay on Bibliographical Control.* (Berkeley: Univ. of California Pr., 1968).
12. Charles H. Davis and Debora Shaw, "Collection Overlap as a Function of Library Size: A Comparison of American and Canadian Public Libraries," *Journal of the American Society for Information Science* 30:19–24 (Jan. 1979).