UC Berkeley

UC Berkeley Previously Published Works

Title

Methodological problems in assessing the overlap between bibliographical files and library holdings

Permalink

https://escholarship.org/uc/item/7hm1431w

Journal

Information Processing & Management, 11(3-4)

ISSN

0020-0271

Authors

Buckland, Michael K
Hindle, Anthony
Walker, Gregory PM

Publication Date

1975-08-01

DOI

10.1016/0306-4573(75)90011-4

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at https://creativecommons.org/licenses/by-nc-nd/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Information Processing & Management, Vol. 11, pp. 89–105. Pergamon Press 1975. Printed in Great Britain

METHODOLOGICAL PROBLEMS IN ASSESSING THE OVERLAP BETWEEN BIBLIOGRAPHICAL FILES AND LIBRARY HOLDINGS

MICHAEL K. BUCKLAND†, ANTHONY HINDLE‡ and GREGORY P. M. WALKER§

Abstract—During 1970–71 the University of Lancaster Library Research Unit carred out a study of the extent to which leading British research libraries tend to duplicate rather than complement each others holdings. This investigation was commissioned in order to provide pertinent background information for the staff of the National Libraries ADP Study[1]. The investigation was in two parts, "National Catalogue Coverage Study" which estimated the overlap in holdings[2] and a "Foreign Books Acquisitions Study" which estimated the extent of duplication in the acquisition of non-British imprints[3]. The authors of this paper collaborated in the design and direction of the investigation.

Several methodological problems were encountered. The purpose of this paper is to identify and discuss methodological difficulties in this specialist type of library survey. Examples of findings from the British study and from subsequent surveys in Indiana are given by way of illustration.

The initial impetus was a major study of the overlap in holdings between 23 British libraries including the leading national and academic libraries. It became apparent that there were severe methodological problems involved.

Since then it has seemed increasingly clear that: (i) When formulated in a general way, overlap is a widely occurring parameter in the bibliographical area whether in automated information processing or manual. (ii) The overlap parameter is likely to often be a critical one from the management point of view, especially in ascertaining the probable costs and benefits of automation, collaboration or, most of all, in collaboration in automation. (iii) Overlap studies are becoming more frequent on account of (i) and (ii).

Therefore this paper has been prepared with the emphasis on the *methodological* aspects rather than the results of either of the actual surveys.

1. OVERLAP AND ITS IMPORTANCE

The basic structure of overlap is represented in Fig. 1. The black dots represent the universe of published items. The circle A represents a library. It encloses those published items which library A holds. Circle B represents another library and encloses those items which library B holds.

There are four sets of items:

(i) Those held by neither library: these are, therefore, outside both circles.

(ii) Those held by *both* libraries: these are, therefore, inside both circles. This "area of overlap" has been shaded.

(iii) Those held by A but not B, in the unshaded part of circle A.

(iv) Those held by B but not A, in the unshaded part of circle B.

Although described here in terms of overlap in library holdings, the overlap structure has much wider application, even in bibliographical matters. In particular, the overlap in the coverage of abstracting and indexing services is fundamentally identical. For example, if each dot in Diagram 1 were held to represent an article on mathematics published in 1969, then circle A could be regarded as enclosing those articles which were abstracted by *Mathematical Reviews* and circle B as enclosing those articles which were abstracted by the *Referativnyi Zhurnal: Mathematika.* The shaded area would, therefore, represent articles abstracted by both. The unshaded area within the circles represents articles covered by one or other service but not by both. The dots outside the circles represents articles covered by neither.

The problem is normally one of assessing the proportion of B which is also held by A. In the notation of conditional probability this can be written as $P(A/B)$, i.e. the probability that an item will be in A, given that it is held by B. The converse is, of course, the proportion of B which is *not* also held by A. In the notation of conditional probability this can be represented as $P(\bar{A}/B)$.

†Assistant Director for Technical Services, Purdue University Libraries, U.S.A.
‡Research Director, Unit for Operational Research in Health Services, University of Lancaster, England.
§Head, Slavonic Department, Bodleian Library, Oxford University, England.

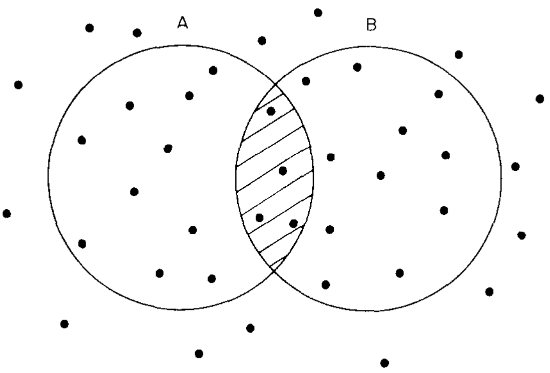

Fig. 1. The Structure of Overlap. The black dots represent the universe of published titles. Circle *A* represents a library. It encloses the titles held by library *A*. Circle *B* represents the holdings of library *B*. The shaded area represents titles held by both.

Since these two conditions are mutually exclusive and no others are possible

$$P(A/B)+P(\bar{A}/B)=1$$

$$\therefore P(A/B)=1-P(\bar{A}/B).$$

The importance of overlap can be emphasized by listing a few examples related to library planning and automation.†

(i) If two or more libraries are considering a collaborative automation programs, then the greater the overlap in holdings the more economical it will be to make use of files held in common.

(ii) The more items research libraries hold in common with, say, the British Museum, then (variant cataloging codes apart) the more they would be likely to benefit if the British Museum's records were converted to machine-readable form.

(iii) The greater the overlap between individual libraries and the national library, then the lower the eventual size of a national union catalog must be and the more nearly the catalog of any one very large library will approximate to the size of such a file.

(iv) A large library with a *low* overlap with the national library would make suitable collaborator if the aim were to achieve broadness of coverage in holdings.

(v) The greater the overlap between an individual library and a cooperative bibliographical center (such as the Ohio College Library Center), the greater the probable benefits of membership in the center in terms of current cataloging or the retrospective conversion of records.

In the case of secondary literature (abstracting and indexing services) similar considerations apply and have been stressed in the SATCOM report[4] and by UNESCO[5]. Instead of the proportion of titles acquired by libraries one is concerned with articles listed by the services.‡

(i) The higher the overlap in coverage between services, the greater the scope for a rationalization and economy through collaboration.

(ii) The lower the overlap between services in the same subject, the greater coverage that could be achieved through collaboration.

(iii) If the interests of a particular user group can be defined in terms of a set of references and

†Although these examples relate to aspects of the National Libraries ADP Study, they were written before the actual recommendations of that study were known and before the library departments of the British Museum were redesignated as the British Library.

‡Occasionally it may prove desirable to compare the journals scanned by the abstracting and indexing services. For an example of this see WOODS *et al.*[6]. However, this is not the same as surveying the overlap of literature indexed because two different services might well scan the *same* journal on a *selective* basis without ever indexing the same article. Indeed, in a later paper the same authors[7] both note that in one year 434 journals were scanned by both CAS (Chemical Abstracts Service) and BIOSIS (the BioSciences Information Service of Biological Abstracts) without any articles in them being indexed by either.

compared with various services, then size of the overlap can be used to compare the relevance of any different services—or groups of services—to that group.†

In brief, overlap can be a key planning parameter in bibliographical matters.

2. EXPERIMENTAL DESIGN

Since it is quite impractical to actually compare two or more libraries *in toto* in any direct way, it is necessary to make partial tests or experiments from which the actual overlap can be estimated. The problem is to develop an approach which will drastically reduce the labor involved and yet produce estimates which are not biased one way or another. A variety of different approaches to bibliographical overlap have been used. They will be examined with respect to economy of effort and likelihood of bias.

2.1 *Comparision of segments of catalogs*

The commonest approach has been to take one or more segments of the author (or name) catalog in each of the libraries concerned and observe how much material was held in common. For example, the letter O has been used in more than one survey.

This approach is easily understood and fairly simple to administer. Nevertheless there are a number of problems.

The most fundamental problem is that some libraries have only partial catalogs. Two of the national libraries of the U.K. are in this category. There is the well-known example of the National Lending Library for Science and Technology (now the British Library Lending Division) which manages without a catalog for most of its stock and the National Library of Wales which does not necessarily catalog fiction material.

Even where reasonably complete catalogs do exist, there are circumstances under which they are not properly comparable on account of variations in cataloging rules. The same item can be entered in different places in different catalogs. Therefore, it cannot be assumed that an item which would be entered under, say, the letter O in one catalog would also be entered under the letter O in another—or that its absence under the letter O in the second library proves that it is not held.

To a large extent this problem can be circumvented by reliance on added entries and cross-references to provide pointers. Also by standardizing on any given set of cataloging rules one can, to some extent, *translate* some headings such as *England* to *G.B.* Nevertheless this problem cannot be entirely solved.

One example which neatly epitomizes this problem is an anonymous pamphlet:

On public libraries: .... Liverpool, 1858.

Suppose that one of the segments chosen for comparison was ONA-ONZ and that the British Museum was one of the libraries being compared. Assuming that AACR rules were taken as the basis for comparison, then this title should clearly be included in the comparison. However, it does not appear in the segment ONA–ONZ of the British Museum's General Catalog of Printed Books. If it appeared in the segment ONA-ONZ of any of the other libraries, then an alert surveyor, bearing in mind the British Museum's rules for the entry of anonymous works, could look for it under *Public Libraries* where it is in fact entered. However, if the British Museum alone held a copy, then it would almost certainly be overlooked unless the surveyor either happened to be aware of this item or was prepared to examine the entire British Museum catalog searching for items which are pertinent to the overlap comparison but which, through the application of local cataloging rules, are entered outside the defined segments. In this example, standardizing on the British Museum rules instead of AACR would not solve this problem unless *all* the libraries being surveyed used British Museum rules. Even then there is some scope for discrepancies. At one library which does use the British Museum rules in order to be compatible, it was ruled that this item would be entered under *Libraries.*

Anonymous works are a particular stumbling block but not the only one. The vagaries of the catalogs of older research libraries should not be lightly dismissed. In one catalog, which is closed but still in use, authors named Smith are entered under F.‡

†For a good example of several management decisions being taken on the basis of an overlap study of secondary literature see ASHMOLE *et al.*[8].

‡For *Fabri*, genitive singular of Latin *Faber*: a smith.

It is certain that variations in cataloging practice will distort overlap comparisons based on segments of the author (or name) catalog. Expert detective work can probably reduce the distortions substantially, but a residual bias of unknown size will remain.

Even if fully consistent catalogs were available, two further problems derive from the variation in the alphabetical distribution of surnames from region to region. Scotsmen, notoriously, have surnames beginning with Mac — and Irishmen with O, while surnames beginning with W are more common in Germany than they are in France. Some data on this is provided by JONES[9].

Because these distributions tend to be reflected in the linguistic and subject composition of a particular library collection, the alphabetical distribution of surnames is likely to vary from library to library.

Furthermore any given segment is likely to represent a linguistic mix which is untypical of the collection as a whole. Because it is untypical, it cannot be assumed that the degree of overlap with respect to that segment is typical of the library as a whole. It should, however, be possible to provide some kind of check on this bias. A prudent course of action would be to take not one segment but several smaller ones on the grounds that the more segments that are taken, the more nearly their individual biases should compensate for each other and the overall effect should approximate to the library as a whole.

It is noteworthy that in a study of the overlap in library holdings in upper New York State by O'Neill, it was considered that a minimum of fifty segments would be needed to reduce bias of this kind to an acceptable level[10].

A more subtle bias can arise from the precise manner in which the segments are picked. Since the alphabetical distribution of authors names depends on the linguistic and, indirectly, the subject interests of a given library, it follows that any defined segment will represent a varying proportion of the whole in different libraries. Conversely, if a set of segments are defined in terms of one library to represent, say, one percent of the whole, then it is quite possible that the same segments would represent less than—or more than—one percent of another library with different interests. Consider the comparison of a small library with predominantly British material with a large library with large holdings of foreign materials. If the segments were defined in the context of the small library, then there is a high probability that the segments will represent parts of the alphabet in which British authors tend to be frequent. The resultant estimate of overlap would probably be overestimated.

If, however, the segments were defined in terms of the larger library, then they are likely to include foreign elements not found in the smaller library and the resultant estimate of overlap would probably be underestimated.

The method of comparing segments of author catalogs appears to have been the most commonly used approach. It is, at least superficially, easy to understand and to handle. Nevertheless, it is certainly not free from bias and was rejected in the case of the overlap studies for the National Libraries ADP Study, on the grounds that a more rigorous approach was required.

Similar considerations apply to the use of segments of a classification scheme. A good example of this approach is the study by ALTMAN[11, 12], who examined the overlap in titles held by public secondary schools and public libraries in two counties in New Jersey.†

Altman identified twelve relatively narrow subject categories known to be relevant to the needs of the users concerned. Each subject was defined in terms of specific numbers in the Dewey Decimal Classification. In each library the titles in these Dewey numbers were noted and compared with the titles drawn from the other libraries. This approach depends on standardised classification practice. Even so, there remains the problem of individually misclassified titles. Titles *wrongly included* in the selected segments can be eliminated by editoral inspection. However, titles *wrongly excluded* are more difficult to identify. In this case: "there was no way to determine how many titles remained unlocated which should have been in the sample, but a detailed analysis of the misclassed (i.e. wrongly included) titles indicates that unlocated (i.e. wrongly excluded) titles would not exceed 4%" (ALTMAN[11, p. 183]).

†Altman's study is noteworthy because it attempts to define user's needs, to clarify what resources are available, to explore the scope for linking needs with resources more effectively and, in particular, to illuminate the dynamics of books selection which results in the overlap patterns observed.

Clearly, the comparison of segments of a classification scheme could only be attempted within a group of libraries which are extremely consistent with respect to classification practice or in which variations in classification practice can be "translated" into a common frame of reference.

2.2 *Sampling from external lists*

A quite different approach is to base the experiment on external lists. For example, if one were interested in overlap in library holdings of, say, modern foreign books then one could pick or sample from foreign national bibliographies. This sample would then be taken to each library and the catalogs searched. The incidence of common holdings would provide the estimate of overlap. This approach has an additional attraction that it can also provide an estimate of the number of items not held by any of the libraries.

Unfortunately bibliographical matters do not lend themselves to this approach because of the difficulties in finding an acceptable sampling frame from which to pick a sample. A genuinely universal bibliography is, of course, the ideal sampling frame, but this does not exist and it is not easy to identify any substitute that would be acceptable for the purposes of studying overlap between libraries.

The situation is much better with respect limited classes of books. For example, one could consider using the leading national bibliographies as sampling frames for picking samples of the modern books from the countries they cover, but the unevenness of the development of national bibliographies means that this cannot be done for all countries; nor, therefore, modern books as a class. This incompleteness is in fact, likely to be a significant source of bias because of the cumulative nature of enumerative bibliography. The inclusion of an item in one list enhances the probability of its being included in additional lists and of its being noticed by library book selectors.

In some cases, this approach is capable of being very laborious. For example, in the case of the acquisition of foreign language books by British libraries, the population of foreign language books is large but the proportion acquired is small. The number of titles acquired by several libraries is smaller still. In order, therefore, to obtain sufficient observations to assess in any detail the degree of duplication of modern foreign imprints it would be necessary to pick a very large sample and indulge in much searching for items not held.

This approach has been used, however, in examining the coverage of abstracting and indexing services. Here, however, the essentially cumulative nature of bibliography can lead to a lack of independence between the source of the sample and the service being tested. For example, one variation of this technique is to identify within the chosen subject area, a bibliography which is relatively comprehensive. This is then used as the sampling frame and some or all of it constitutes the sample which is checked against abstracting and indexing services which purport to cover that subject. This can result in the use of a bibliography as a sampling frame for the testing of abstracting and indexing services which was itself compiled from those services[13]. Such results are likely to be biased unless the bibliography concerned is entirely comprehensive or unless the bibliography was compiled without using abstracting or indexing services. In the latter case, membership lists of a professional society can be used[14]. However, a bibliography compiled in this manner would be likely to be incomplete representation of the world-wide literature on any subject—or even of the literature in any given language.

The technique of sampling from external lists suffers from the incompleteness of available lists and a lack of independence between the test sample and the whatever is being tested. It can also be laborious.

2.3 *Direct statistical approach*

It is possible to avoid the effects of variant cataloging rules and reliance on external lists by a more direct statistical approach. This is done by picking a random sample from A (e.g. a library or an indexing service) and checking this sample against the holdings of B (e.g. another library of indexing service). The proportion of the sample which is held by B is taken as the estimate of the overall proportion of A which is also held by B. For examples of this see Tables 1 and 2. The actual number of items held in common is assessed by multiplying the number of items in A by the proportion estimated to be also held by B. More generally B can represent a group of libraries.

The procedure is reversible in that, as an independent check, it is possible to pick a sample from B and check it against the holdings of A.

It is assumed that the number of items found in the sample of items from A also found in B is a binomial random variable. Further it can be shown that under certain circumstances this proportional overlap is approximately normally distributed. As a rule of thumb this approximation can be made if n is large enough to make $n\pi > 5$ and $n(1-\pi) > 5$ where

$n =$ the number of items sampled from A and

$\pi =$ the proportion of the sample from A also found in B.

If these conditions do not hold it is likely that the Poisson distribution will give a better approximation than the normal.

If the Normal distribution is used the confidence intervals for π can be obtained from

$$\pi = P \pm Z \, . \sqrt{\frac{P(1-P)}{n}}$$

where $P =$ the observed proportionate overlap

$Z =$ the value of the standard normal variate

Table 1. Paired overlaps: British research libraries. This table should be read: An estimate based on 942 observations indicates that 37% of the pre-1968 monographs in the British Museum are also held by Cambridge University Library. (Data extracted from University of Lancaster[2], p. 14.)

	Observations	British Museum	Cambridge U. L.	Bodleian, Oxford	Nat. Libr. Scotland
British Museum	942	-	37%	41%	30%
Cambridge U.L.	231	78%	-	73%	58%
Bodleian, Oxford	239	69%	57%	-	54%
Nat. Libr. Scotland	236	81%	74%	77%	-

Table 2. Paired overlaps: Current cataloging in selected Indiana libraries versus Ohio College Library Center data base. This table should be read: An estimate based on 175 observations indicates that 62% of the acquisitions of the Indiana State Library were established in the OCLC data base when processed at the Indiana State Library. (Data from MARKUSON[15], Tables B-6 and B-7, pp. A85–86)

			Titles found in OCLC data base	
Library	Observations	MARC	Shared	Total
Indiana State Library	175	49%	13%	62%
Indiana Univ. Libraries				
- Bloomington Campus	60	43%	7%	50%
- Regional Campuses	60	53%	32%	85%
Purdue Univ. Libraries (West Lafayette Campus)				
- Current orders	178	49%	15%	64%
- Current cataloging	150	61%	27%	88%
Detroit Diesel Div. GMC, Library	20	80%	20%	100%
U. S. Naval Avionics Library, Indianapolis	20	95%	5%	100%
Indianapolis-Marion -County Public Library	270	70%	15%	85%
Anderson Public Library	26	42%	38%	81%
Duneland School Corp.	50	70%	16%	86%
Indianapolis-Perry Township School Corporation	80	31%	55%	86%

For the 95% confidence limits $Z = 1{\cdot}96$. If the Poisson distribution is used the distribution of P is asymmetric. The upper confidence limit is given by

$$\frac{1}{2n}\cdot \chi_\alpha^2$$

where $\alpha =$ the required confidence level
(i.e. $\alpha = 0{\cdot}05$ for the 95% confidence limit).
In other words the distribution of P is approximated by the chi-squared distribution.

The asymmetrical characteristic can be important if the number of unique items in A is being assessed where the level of overlap between A and B is close to unity. In these circumstances there can be significant error in the direction of over-estimating the number of unique items. Similarly, if the level of overlap is close to zero, there can be significant error in the direction of under-estimating the number of unique items. However it is important to note that in all circumstances the observed value P is an unbiased estimate of π.

In the case of overlap studies involving more than two libraries, this approach can be extended by picking a sample from each library and checking it against each other library. With careful experimental design, this approach can be used for a wide variety of analyses. The overlap can be estimated not only between any one library and any other single library but also between any one library and any group of two or more of the other libraries. In each of these cases, the structure of the problem is essentially in the form of a paired overlap which can be represented as

$P(N/A)$—the probability that an item is in N given that it was sampled from A, where N can represent one or a group of libraries.

Overlaps involving *groups* of libraries require careful handling, as will be illustrated by the following notes on four distinct problems.

Group problem 1: *Duplication within a group.* Consider a group of libraries: To what extent is the material they acquire duplicated within that group? How much of the material is held by one only, by two,... by all? This type of inquiry (the essence of the Foreign Books Acquisition Study) can be tackled by taking a proportionate sample from each of the libraries in the group (e.g. a 1 in 500 sample from each). In each case each item is checked against the holdings of each other library, and the number of items held by 1, 2, 3,... $\underline{n}$ libraries noted.

These results are, however, biased by the sampling technique used and need to be adjusted. This bias arises in the following manner. If a title were only held in one library, then it would have had a 1 in 500 chance of being sampled. In other words, the probability of *not* being sampled would be 0·998 for an item held by one library but 0·998 × 0·998 (approx. 0·996) for an item held by two libraries. More generally the probability that an item would be included in the sample if it were held by $\underline{n}$ libraries is $1\text{–}0{\cdot}998^n$. In brief, this approach seriously overestimates the amount of duplication due to the increase of sampling density with increased duplication. Duplicated titles are more likely to appear in the sample than would unduplicated titles. The observations of sampling density need, therefore, to be weighted by some function of the sampling density—in this case $1/(1\text{–}0{\cdot}998^n)$. For an example see Table 3.

Tackling this problem with non-proportionate samples is not recommended.

Group problem 2: *Duplication of a given library's holdings.* Given that a book is held by library A, what is the probability that it is also held in none, one, two, . . . , all of libraries B, C, D and E? If an article is abstracted in *Mathematical Reviews*, is it likely to be listed in none, one, two or all of the *Referativnyi Zhurnal*: *Matematika*, the *Zentralblatt für Mathematik und ihre Grenzgebiete* and the *Bulletin signalétique*: *Mathétiques pures et appliquées*? This type of duplication can be estimated by picking a sample from library A or *Mathematical Reviews* and checking the item in the sample against the other files.

Since one is concerned with only one library's holdings at a time it does not matter whether the samples from the various libraries are proportionate or not. For an example see Table 4.

Group problem 3: *Cumulation into groups.* How would a file grow if collections were added serially? An example of this is the construction of an union catalog or bibliographical data base by selecting one library then choosing others to add to it. An example is given in Table 5. Another example would be a library which possessed one indexing or abstracting service and was interested in the marginal increase in coverage that addition services might provide.

Table 3. Duplication within a group: non-British monographs dated 1950–1967 held by a group of eighteen British research libraries. This table should be read: 49% of titles from the U.S.A. and Canada were held in just one of the eighteen libraries and 13% were held in two libraries. The total number of titles from this area is estimated as 168,900 titles held in 491,900 copies, giving a mean level of duplication of 2·9 copies per title. (Data extracted from University of Lancaster[3], p. 10.)

	Political region			
No. of Libraries holding a copy	U. S. A. & Canada	Western Europe	Eastern Europe	Elsewhere
1	49%	56%	65%	60%
2	13%	19%	21%	22%
3	8%	11%	8%	9%
4	7%	6%	3%	4%
5	6%	4%	1%	2%
6	6%	2%	1%	1%
7	3%	1%	<0.5%	1%
Approximate numbers involved				
- titles	168,900	65,500	129,400	252,600
- copies	491,900	114,400	206,500	500,900
Mean level of duplication	2.9	2.0	1.6	1.7

Table 4. Duplication of a given library's holdings: Pre-1968 monographs held in a group of 18 British research libraries. This table should be read: An estimate based on 942 observations indicates that 42% of the pre-1968 monographs in the British Museum library are held in none of the other 17 libraries in the group—and 16% are held by just once in the other libraries. (Data extracted from University of Lancaster[2], p. 20.)

	Observations	% of holdings also held by 0, 1, 2, 3, 4 or 5 of the other libraries in the group of 18 British Research Libraries.					
Holding Library		0	1	2	3	4	5
British Museum	942	42%	16%	10%	11%	10%	4%
Cambridge U. L.	231	10%	10%	13%	22%	20%	10%
Bodleian, Oxford	239	19%	15%	11%	21%	11%	8%
Nat. Lib. Scotland	236	6%	7%	11%	28%	22%	7%

Table 5. Cumulation into groups: Extending the British Museum catalog. This table should be read: It is estimated that the catalog of the British Museum Library would be increased by 450,000 titles if the pre-1968 monographs of the Bodleian Library, Oxford were added to it or by 315,000 if the holdings of Cambridge University Library were added; or by 670,000 if both were added. (Data extracted from University of Lancaster[2], p. 15.)

Addition of one library only	Estimated contributions
Bodleian, Oxford	450,000
Cambridge U. L.	315,000
Nat. Lib. Scotland	218,000
Nat. Lib. Wales	156,000
Addition of two libraries only	
Bodleian & Cambridge U.L.	670,000
Nat. Lib. Scotland & Nat. Lib. Wales	333,000
Addition of three libraries only	
Bodleian, Nat. Lib. Scotland and Nat. Lib. Wales	680,000
Addition of four libraries only	
Bodleian, Cambridge U.L., Nat. Lib. Scotland and Nat. Lib. Wales	872,000

Let us consider the cumulation of titles as five libraries A, B, C, D and E are added, in that order.

(i) *Proportionate samples.* If the samples drawn were proportionate then one would proceed in three stages:

—consider A alone;

—use a paired overlap to assess the contribution derived by adding B;

—perform a series of analyses in the manner of Group Problem 1 above for each successive group ($A \cup B \cup C$; $A \cup B \cup C \cup D$; $A \cup B \cup C \cup D \cup E$) in order to estimate the number of different titles contained in each grouping and the increase resulting from each successive increment in the size of the groups of libraries.

(ii) *Non-proportionate samples.* If the samples drawn were not proportionate then estimates can still be made if the cumulation is seen as a series of paired comparisons in the form $P(N/M)$ where N is the existing group of libraries and M is the next library to be added. More specifically one would cumulate five libraries in the following manner which is depicted graphically in Fig. 2.

—consider A alone;

—use a pared overlap $P(\bar{A}/\bar{B})$ to assess the contribution derived by adding B;

—use a paired overlap in the $P(\bar{A} \cup \bar{B}/C)$ to assess amount held by C which is not in A or B;

—similarly use $P(\bar{A} \cup \bar{B} \cup \bar{C}/D)$ and $P(\bar{A} \cup \bar{B} \cup \bar{C} \cup \bar{D}/E)$ to estimate the contributions of D and E respectively.

Naturally one might well want to compare the effects of alternative sequences, e.g. A, D, C, B, E compared with A, B, C, D, E. If one sought to increase the range of titles in the union catalog at least cost, then, if the unit cost of adding items does not change, the best strategy is to add as the next library, that which has the lowest proportionate overlap with the existing file regardless of either the size of the file or the size of the library. This is in keeping with search theory and has some independent support from empirical analyses by ARMS[16].

Group problem 4: *Comparison of groups.* How many different titles are there in the group D, E and F which are also held in the group A, B and C?

(i) *Proportionate samples.* It is convenient to take the three proportionate samples from D, E and F as being representative of the group and check them against A, B and C and observe the match. However, to be rigorous one should need first to check the samples from D, E and F and check each of them against the other libraries in the same group in order to establish duplication within the group in order to allow for variant sampling density within D, E and F and weight them accordingly as described above.

(ii) *Non-proportionate samples.* Comparing groups of libraries when non-proportionate

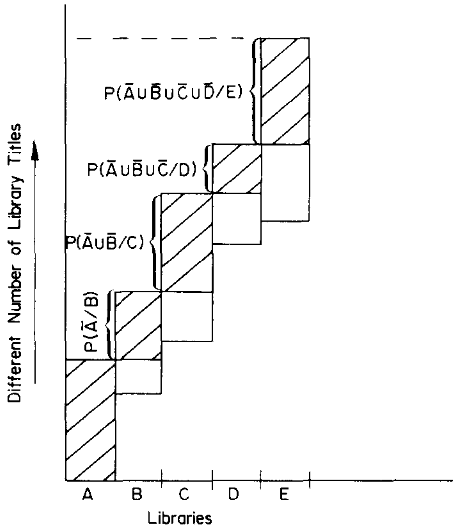

Fig. 2. The cumulation of titles as the number of libraries in the group is increased. The shaded areas represent the titles contributed by each library.

samples have been drawn appeared to be a major stumbling block in the Lancaster overlap studies.

The approach eventually adopted was an extension of that used for non-proportionate samples in group problem 2: Cumulation into groups. This involved using that technique to estimate separately the number of titles cumulated in the three groups.

(i)—*D*, *E* and *F*

(ii)—*A*, *B* and *C*†

(iii)—*A*, *B*, *C*, *D*, *E* and *F*.

The answer to the question "How many of the titles in *D*, *E* and *F* are *not* in the group *A*, *B* and *C*?" is given by the difference between (ii) and (iii). The answer to the question "How many of the titles in *D*, *E* and *F* are also in the group *A*, *B* and *C*?" is given by (i) minus the difference between (iii) and (ii). The value of (i) can also be used to calculate either quantity as a proportion of *D*, *E* and *F*.

It may be noted that proportionate samples become unmanageable where the libraries vary substantially in size. With a large proportion, the amount of data sampled from the largest libraries become unmanageably large. With small proportions, the number of observations from the smallest libraries become trivial and, in consequence, the reliability is low.

2.4 *Other experimental designs*

SCHOFIELD and URQUHART[17] have reported an ingenious approach designed to estimate the *eventual* overlap in current material. If we have understood their approach correctly, it has three stages:

(i) All orders are noted from each of the libraries for a fixed period.

(ii) The degree of duplication within that original period is noted, i.e. some proportion of the titles ordered by library *A* will also have been ordered by library *B* within that initial period.

(iii) An attempt is then made to estimate the eventual overlap for this material. In other words, some of the items ordered by library *A* during the initial period, may be ordered by library *B* *after* the initial period. Therefore, the degree of duplication can be expected to rise with time. Statistical calculations are made to estimate the *limit* to which this duplication is likely to rise.

The results obtained amongst the libraries of London University indicated a surprisingly small amount of duplication. A useful feature of this technique is that its results can be verified by checking the original sample from *A* against the catalog of *B* after a year or two.

NUGENT[18] has presented a correctly formulated statistical approach which he used to assess the overlap in collections between the libraries of six New England state universities. PARKER[19] subsequently adopted Nugent's approach but, as an economy, took a single undifferentiated sample from the non-proportionate samples which had been taken from each of the five libraries which he studied. This economy was, however, statistically improper, for it is not clear what the sample is supposed to represent.

3. EXTRAPOLATION

The marginal contribution achieved by adding a library to a group, as when compiling a union catalog, can be assessed rather directly by means of overlap estimates as has been described above, so long as the number of libraries is small and tractable.

However it is tempting to explore the feasibility of making projections beyond a small group in order to estimate the marginal contributions made where large groups of libraries are concerned or the total number of titles held by a large group of libraries—even the total number of titles in all the libraries of a country. Such extrapolation is, of course, rash, nevertheless it does seem worthwhile to attempt some tentative estimates in the absence of any other available data.

This problem was tackled in the Lancaster survey and has subsequently been approached from a different standpoint by ARMS[14] using the Lancaster data.

The approach used in the Lancaster survey involved the use of two models. The first model was based on a Bradford–Zipf relationship of the form

$$m(r) = \frac{k}{(r+c)^{a}}$$

†This figure is produced in the course of generating (iii).

where $m(r)$ is the marginal contribution of the rth library, r is the rank order of the library (in size) and c, k and θ are constants. The appeal of this model lies in the fact that the Lancaster survey obtained data on the largest eight libraries in the United Kingdom and the model fits these data well. Extrapolation is always subject to error but the model was used to determine a "lower-bound" figure for the number of different monographs in the estimated total of 5,000 British libraries in the United Kingdom. The estimate is a lower-bound one because the eight libraries are large general libraries (five of them with copyright deposit privileges) and it is to be expected that the addition of other, more specialized libraries will yield a higher proportion of unique items than that predicted by the model. In other words, if these large general libraries are assumed to be atypically homogeneous, the marginal productivity is, thereby, assumed to be atypically low. Therefore an extrapolation based on them is assumed to underestimate total number of titles.

The second model was derived from a plot of the cumulative number of unique titles for the eighteen libraries and the discovery of a linear relationship for libraries ranked fifth to eighteenth (by size) on semi-logarithmic paper. This estimate was regarded as an "upper-bound" estimate because several of these libraries were chosen because of their specialized properties and would yield a "higher than average" proportion of titles than would a less biassed group of libraries and, therefore, an exaggerated estimate of the total number of titles.

These two models gave estimates of 4·15 and 5·4 million titles, respectively, for the number of unique titles in the libraries of the United Kingdom.

Arms[16], using the Lancaster data, examined the number of books held by exactly one, two, through to eleven or more libraries and derived the ratio

$$f(i)/f(i+1)$$

where $f(i)$ is the number of books held by exactly i libraries. This ratio suggested the model

$$f(i+1)=\theta f(i)$$

or equivalently

$$f(i)=a\theta^{i-1}(0<\theta<1)$$

where $f(i)$ is a constant between zero and 1. In other words this ratio appeared, surprisingly, to be relatively stable and was treated as a constant. Arms also demonstrates that this sample model fits the number of locations per item in the Union Catalogue of Books in the National Central Library, London, now part of the British Library Lending Division.

From this model Arms derives an expression for the total number of different titles in a collection of libraries, viz.

$$K=a(1-\theta^{M})/(1-\theta)$$

where K is the number of unique titles and M is the number of libraries. He further demonstrates that for large values of M the number of extra titles added by adding a library having n books becomes

$$n(1-\theta).$$

It is clear that in this model θ is a fundamental parameter and, despite its virtue of simplicity, more evidence is required to support the model particularly when large groups of libraries are being considered. Intuitively it seems unlikely to be true that the marginal contribution of a library will be the same when added to a group of five other libraries as when added to a group of five hundred other libraries. In general the parameter can be interpreted as a measure of the homogeneity of the group of libraries and where $\theta=1$ libraries are identical, where $\theta=$ zero the libraries have nothing in common.

4. NOTES ON SAMPLING, EDITING AND CHECKING

The direct statistical approach described in Section 2.3 above presupposes that a random sample was taken from each library. Unlike the approach based on the comparison of catalog segments, it does not presuppose cataloging rules are basically consistent from library to library—or even that a catalog exists. It is only necessary that a sample can be picked. This gives considerable flexibility in adapting sampling techniques to local circumstances.

Sampling from library collections has several pitfalls, especially if the surveyor attempts to sample directly from the shelves since the number of volumes per shelf may well vary in a biased fashion between subject or age groups and the composition of the array of books actually on the shelves at any given point in time is invariably biased towards the less used materials.†

The most satisfactory solution seems to be to sample from a list of the libraries holdings and, as is normal practice in sampling from lists, pick a starting point at random and then take every *n*th item, continuing through the end of the list into the beginning and until the original starting point is reached. In practice this involves severe practical problems.

(a) An ideal list would give each title once and once only. In other words, a title which is held in two copies ought not to have an enhanced probability of being sampled. In smaller and newer libraries there is often a shelf-list either in this form or such that duplicated titles are self-evident and can be discounted. Otherwise the author catalog can be used, but there is the problem that whereas the sampling of cards is the only practical approach, the existence of entries for second authors, editors, etc., ensures that the number of cards per title varies. Since modern cataloging practice is based on the principle of having a single recognisable main entry per title and a variable number of added entries it is possible to examine every *n*th card but to ignore cards which are not main entries. Whatever the mean and variation in number of cards per title may be, this procedure will isolate approximately one *n*th of the main entries represented in the catalog and those sampled will have been taken more or less evenly from the whole catalog. The Birmingham (England) Public Reference Library was excluded from the National Catalog Coverage Study because its catalog of pre-1879 holdings contained no distinction between main and other entries. Since the number of titles in this catalog (which was being revised) was not known, it was impossible to pick a sample with any confidence that it would in fact constitute the desired proportion of the whole population of titles represented.

(b) A practical problem is that even with a medium-sized library, life is too short to permit the rigorous sampling of *precisely* every *n*th card. However, this problem can be very considerably reduced by adopting a two-tier approach. For example if a one-five-hundredth sample is desired then one could, amongst numerous other strategies, measure one inch in every fifty as an initial subset and then examine every tenth card in these measured inches. In the case of catalogs in book form, it is sensible to examine every *n*th page. Naturally this "telescoping" procedure could, in itself, be a source of bias if the number of measured inches or book pages were small. In the National Catalog Coverage Study a lower limit of 200 such points was, rather arbitrarily, regarded as minimal.

(c) Linear measurement of card files is susceptible to at least two sources of bias.

(i) The number of cards per inch depends on the tension with which they are held. This especially noticeable with heavily used catalogs. Short of special equipment the best that can be done is for one person only, who is conscious of the problem, to do all the holding of the cards during measurement. This makes for consistency.

(ii) In most libraries the thickness of the card stock has varied over the years. For this reason when a pin is inserted at a measured point, the *next* card should be taken even if the pin actually sticks into a card. Otherwise there will be an enhanced probability of sampling relatively thick cards relating to a particular cataloging period. One catalog study is said to have produced distorted results from this cause.‡

Naturally the problems of sampling are substantially reduced if the number of cards in the catalog or the number of titles in the library are reliably known in advance. This is, unfortunately, rare.

Having established a sample there is the problem of editing it to exclude material outside the scope of the study. In the case of the National Catalog Coverage Study, there were three classes of exclusions.

(i) Everything published after 1967 was excluded. This reflects a dilemma which deserves some comment. The decision to exclude imprints dated 1968 or later was based on the primary

†For a discussion of "collection bias" see BUCKLAND *et al.*[20] or BUCKLAND[21].

‡Even if the story is apocryphal it does illustrate the variety of the pitfalls to be avoided.

emphasis of the survey—on duplication in holdings. Since selection procedures, ordering routines, book suppliers and cataloging all involve delays, an appropriate interval had to be left to permit records of material to reach the catalogs. Strictly this problem is insuperable because libraries can and do acquire material published centuries ago. Nevertheless a library's holdings of the imprints of a given year can be expected to stabilize after a couple of years. The end of 1967 was taken as a compromise between up-to-dateness and reliability.

(ii) Materials not in a European language were excluded.

(iii) Non-monographic material was excluded as outside the scope of the study. Whilst newspapers, maps, prints, typescript theses and other obvious materials posed no problem, there always remains a penumbra of uncertainty. In these circumstances the best that can be done is to retain the services of a bibliographer experienced in foreign bibliography and, after various trials and discussions to establish case law, leaving all decisions to one person in order to maintain consistency.

A related bibliographical problem emerges when the samples are being checked against the holdings of other libraries. In the National Catalog Coverage Study, this checking was normally done by members of the cataloging staff of the library concerned. Each checker was instructed: "For each item, ask yourself the question 'Has our library this edition?'" The survey was handled at edition level because the alternatives seemed worse. Certainly, of the pursuit of variant issues there is no end. Occasionally there is genuine cause for uncertainty as to whether two documents are of the same edition or not. This problem is nicely epitomized by the following statement while recently appeared in a professional library journal: "*Library surveys*, by Maurice B. Line . . . has been reprinted with substantial amendments and additions which, however, do not in the opinion of the publishers . . . quite justify the reprinting described as a new edition"[22].

Mercifully this dilemma presented itself infrequently.

A final problem in this area relates to the thoroughness of the checking itself. There ought, of course, to be no doubt, given an entry from the catalog of a major research library whether or not the item concerned is or is not represented in the catalog of another major research library—especially if the check is performed by someone familiar with the catalog in which the search is being made. Nevertheless it seems reasonable to believe that in any approach involving checking the most common source of error in the results will stem from the occasional failure to find a book during the checking which was nevertheless held. It follows that the estimates of overlap will tend to err on the side of *underestimating* the overlap. This would have the effect of overestimating the number of different items in a group of libraries. Insofar as the results may be used to justify policies on the basis of an assumed degree of overlap, any error of this kind can be regarded as a safety margin in that there is *at least* this degree of overlap.

It is, of course, possible to cross-check the checking to verify or ensure its reliability. This was done on three occasions during the National Catalog Coverage Study. An expert scrutiny of part of the checking carried out at the National Library of Scotland revealed a high standard of reliability. Another scrutiny, however, at a library of less central interest (and uncommon cataloging rules), revealed lower reliability. In the case of the British Museum which was clearly pivotal to the study, almost all of the items not found during the first checking were re-checked by the British Museum's own staff.

5. DATA HANDLING

There are rather severe technical problems in handling data in an overlap study of any complexity. In the Lancaster overlap study data were handled in two stages: checking and statistical analysis.

5.1 *Checking*

The sample of titles excerpted from catalogs, etc., was reproduced by photocopying or transcribing the original entries. Each entry was assigned the location code of the library from which it was sampled and was edited bibliographically, if needed, to make its entry conform to standard cataloging practice. The entire sample (some 23,000 items) was then alphabetised, mounted on sheets of paper and numbered. Since, in most cases, only a subset of the original proportionate samples from each library was to be checked, these "sub samples" had to be picked from the list and the status of the individual titles marked—mainly by putting a bar across those not to be checked. Several copies of the sample were then produced by photo-offsetlithography and copies were sent out to each library for the items to be checked against their holdings (excluding, or course, items sampled from that library). Items found were marked with the location code of library in which they were found and any queries of bibliographical identity were resolved. The marked-up lists returned from each library were then collated with a master copy of the list and the additional locations noted. Meanwhile each title was coded by date of publication, place of publication and language in order to permit more detailed analyses.

5.2 *Statistical analysis*

Mr. Mel Dobson, then a scientific programmer at the University of Lancaster, now of Preston, England was primarily responsible for the design and development of a suite of computer programs which proved both convenient and effective.†

5.2.1 *Data file.* The format of the data file is best described as a large matrix with one row for each title containing fixed fields containing information about that title.

The first field contained the identification number of the title concerned. This provided a link to the photocopied details on the list of samples.

The next three fields contained bibliographical data concerning the title:

—language:

—country of publication;

—year of publication.

The final field is an array, with one position for each of the libraries in the survey. The value 1 was entered in the position corresponding to the library from which the title was sampled. Note that it is possible, though unlikely, that a given title could, by chance, be sampled from more than one library. The value 2 was entered in the positions which corresponded to libraries in which that title was checked and found to be held. The value O was entered in all other positions, signifying that the title was neither sampled from nor checked and found to be also held in those libraries. Table 6 illustrates the nature of the matrix.

It may be noted that if the survey were to be extended at a later date additional libraries, sample titles and checking data can be added in by expanding the matrix and entering the appropriate codes.

In the Lancaster overlap study there had to be three files:

—The entire file of proportionate samples, not all of which was checked. This was used only for counting purposes [(i) below].

—A file of the proportionate samples of modern foreign titles, all of which were checked. For convenience, the data matrix of the entire file was used with place of publication and date of publication restrictions to exclude irrelevant titles.

—A separate file of the non-proportionate sub-samples which were checked for the overall overlap estimates. This was a subset extracted from entire file and analysed separately.

5.2.2 *Overlap analyses.* A suite of programs was developed to "interrogate" the data base.

(i) *Counting.* One enquiry was a straight count of the sample size of a given library. This was done by scanning the entire data base, title by title, and counting the number of times a 1 occurred in the position corresponding to the library concerned.

Table 6. Data Matrix. This may be read as follows. The item 1 which was published in Russia, in Russian, in 1899 was sampled from library *C* and is not held in any of the other libraries. Item 2, which was published in Switzerland in French in 1735, was sampled from library *A* and is also held only by library *E*. Item 3, which was published in England, in English, in 1965, was sampled from both library *C* and library *E* and is also held by libraries *F* and *G*

I.D. No.	Language	Country	Year	Libraries						
				A	B	C	D	E	F	G
00001	SR	ER	1899	0	0	1	0	0	0	0
00002	RF	WS	1735	1	0	0	0	2	0	0
00003	GE	WE	1965	0	0	1	0	1	2	2

†Enquiries concerning the programs should be addressed to the Library Research Unit, University of Lancaster, Lancaster, England.

(ii) *Overlap between pairs of libraries.* The overlap $P(A/B)$ is estimated by scanning the entire file, counting:

(a) the number of titles with a 1 under library B;

(b) the number of titles in a which also have a 2 or 1 under library A. Those with a 1 under library A are included because if they were sampled from library A it is self-evident without checking that it is held by library A.

The probability $P(A/B)$ is given by dividing count b by count a. Furthermore the count a, being the sample size, may be used to compute the estimated reliability of the answer by reference to appropriate statistical tables.

(iii) *Marginal productivity.* An extension of the paired overlap is the case of $P(N/A)$ where N represents a group of libraries. How many titles would be added to the union catalog of libraries X, Y and Z if library A were added? This is formulated in the same way as the paired overlap [(ii) above] except that count b is re-defined as the number of titles in a which also have a 2 or a 1 under one or more of libraries X, Y and Z.

The programs developed by Mr. Dobson contained the convenient feature of accepting a list of libraries in any given order and then generating at one pass of the data base a series of marginal productivity analyses showing the proportion of each libraries titles which would be new to a union catalog if they were added in the order specified. To estimate the marginal productivity in terms of the absolute number of titles added, of course, each proportion would have to be multiplied by the estimated number of titles held by each library.

(iv) *Duplication of a given libraries holdings* (*Group problem No.* 2). Given that a title is held by Library A, what is the probability that it is held by 0, 1, 2, 3, . . . , all of a specified list of libraries? To do this the program must search the file for titles with a 1 under library A, then count how many times for that title there is a 1 or a 2 in the positions corresponding to libraries specified in the list.

(v) *Duplication within a group* (*Group problem No.* 1). Considering a group of libraries, to what extent do titles tend to be held by only one of the group, by two, . . . by all. If proportionate samples are used, an estimate can be produced in the following manner. Specify a group of libraries. Search the data matrix for titles with a 1 corresponding to any one or more of the specified libraries. In each such case count the number of libraries with a 1 or a 2. The distribution of such counts would need to be weighted by sampling density (as described above) but would then answer the question posed. With nonproportionate samples this approach would be improper, but some indications would be provided by a series of analyses (one for each of the libraries concerned) of duplication of a given libraries' holdings [(iv) above].

(vi) *Sub-analyses.* Coding by language, place of publication and date of publication permits enquiries to be based on a wide range or subsets. For example, all books in French, all books published in Ireland, or all books published between 1918 and 1939.

Combinations of these can be performed. For example, one might enquire what proportion of the titles in Gaelic, published outside Scotland, before 1900 and held in the National Library of Scotland are also held in Cambridge University Library. However, the more exotic the combination the more likely it is that the number of observations will be too small to permit any confidence to be placed in the reliability of the estimate.

(vii) *Lists.* A convenient facility is the option of being able to print out a transcript of the data matrix of all titles relevant to a given enquiry. This can be used to verify unlikely results because the identification number of the title relates to the full bibliographical description as sampled and the library codes indicate which libraries as supposed to hold it. One further possible use would be to list out titles identified as being above a given level of duplication within a group. This could save much time in the collection of data comparing the chronology of ordering or receipt of items—or of the timeliness of the reporting of articles in abstracting and indexing services.†

6. SUMMARY

Bibliographical overlap, whether it be in the holdings of libraries or in the coverage of abstracting or indexing services, is a parameter which is significant in several ways for planning purposes. It is relevant to library selection policies, library processing (especially cooperative

†The study by ASHMOLE *et al.*[8] is interesting for its attention to timeliness.

computer-based cataloging) and to the bibliographical control of subject literatures. It is to be expected that overlap studies will become increasingly common.

The methodological problems are complex and derive from inconsistencies in cataloging practice, the inherently cumulative nature of bibliography and the need for reliable sampling and careful experiment design.

The authors have drawn on their experience of a major overlap study and have analysed a variety of alternatives in the formulation of overlap experiments. Examples are given by way of illustration. Finally a powerful approach to handling the experimental data is presented.

Acknowledgements—The authors wish to acknowledge the help of Mr. A. Graham Mackenzie, University Librarian, University of Lancaster, Mr. Ian M. Stuart, Assistant Librarian, University of Lancaster Library, and Mr. Maurice B. Line now Director, British Library Lending Division and Dr. John Hawgood, Durham University. Mr. Line and Dr. Hawgood were respectively Director and Consultant on the National Libraries ADP Study for whom the Lancaster overlap studies were performed.

REFERENCES

[1] The results of this study are reported in: GREAT BRITAIN. Department of education and science. *The scope for automatic data processing in the British Library.* London, Her Majesty's Stationery Office, 1972. 2 v.

[2] UNIVERSITY OF LANCASTER. Library research unit. *National Catalogue coverage study. Report to the National Libraries ADP Study.* Lancaster. University of Lancaster Library Research Unit, 1971. *Available from the British Library Lending Division, Boston Spa, Yorkshire, England LS23 7BQ as*: NAB 800-N The scope for automatic data processing in the British Library. Supporting Paper N: National catalogue coverage study. Library Research Unit, University of Lancaster.

[3] UNIVERSITY OF LANCASTER. Library research unit. *Foreign books acquisition study. Report to the National Libraries ADP Study.* Lancaster, University of Lancaster Library Research Unit, 1971. *Available from the British Library Lending Division, Boston Spa, Yorkshire, England LS23 7BQ as*: NAB 800-L The scope for automatic data processing in the British Library. Supporting Paper L: Foreign books acquisitions study, Library Research Unit, University of Lancaster.

[4] COMMITTEE ON SCIENTIFIC AND TECHNICAL COMMUNICATION (SATCOM). *Scientific and technical communications: a pressing national problem and recommendations for its solution.* National Academy of Sciences, Washington, D.C. 1969. (pp. 146 & 174).

[5] UNESCO. *Survey of the organization and functioning of abstracting and indexing services in the various branches of science and technology.* Paris. UNESCO, 1962.

[6] J. L. Woods, C. Flanagan and H. E. Kennedy: Overlap in the lists of journals monitored by BIOSIS, CAS and Ei. *J. Am. Soc. Inform. Sci.* 1972. **23**(1), 36–38.

[7] J. L. Woods, C. Flannagan and H. E. Kennedy: Overlap among the journal articles selected for coverage by BIOSIS, CAS and Ei. *J. Am. Soc. Inform. Sci.* 1973, **24**(1), 25–28.

[8] R. F. Ashmole, D. E. Smith and B. T. Stern: Cost effectiveness of current awareness sources in the pharmaceutical industry. *J. Am. Soc. Inform. Sci.* 1973, **24**(1), 29–39.

[9] K. P. Jones: Subcodes: an examination of their utility with special reference to feature card systems. *Aslib Proceedings* 1971, **23** (5) 237–246. Note that Jones' data are derived principally from telephone directories. It should be noted, however, that the tendency for works of joint authorship to have the names of the authors cited in alphabetical order on the title page can be expected to lead to a distribution of main entries in a library catalogue which is systematically biased toward the beginning of the alphabet compared with distributions in telephone directories.

[10] Private communication from Professor E. T. O'Neill. For details of the survey concerned see: E. T. O'Neill. *A survey of library resources in western New York.* Buffalo, N.Y., State University of New York, School of Information and Library Studies, 1971.

[11] E. O. Altman: Implications of title diversity and collection overlap for interlibrary loan among secondary schools. *Lib. Quart.* 1972, **42**(2), 177–194.

[12] E. O. Altman: *The resource capacity of public secondary school libraries to support interlibrary loan: a systems approach to title diversity and collection overlap.* Ph.D. thesis. Rutgers university, 1971. (University Microfilms order no. 71-20,040).

[13] W. E. Bost: Tests on abstracts journals. *J. Docum.* 1968, **24**(1), 61.

[14] R. R. Montgomery: An indexing coverage study of toxicology literatur. *J. Chem. Docum.* 1973, **13**(1), 41–44.

[15] B. E. Markuson: *The Indiana Cooperative Library Services Authority—A plan for the future. Final project report of the Cooperative Bibliographical Center for Indiana Libraries (COBICIL) Feasibility Study.* Indianapolis, Indiana State Library, 1974.

[16] W. Y. Arms: Duplication in union catalogs. *J. Docum.* 1973, **29**(4), 373–379.

[17] J. A. Urquhart and J. L. Schofield: Overlap of acquisitions in the University of London Libraries. *J. Librarianship* 1972, **4**(1), 32–47.

[18] W. R. Nugent: Statistics of collection overlap at the libraries of the six New England state universities. *Library Resources and Technical Services* 1968, **12**(1), 31–36.

[19] R. Parker: *A feasibility study for a joint computer center for five Washington, D.C. university libraries; final report.* [Washington, D.C.] Consortium of Universities of Metropolitan Washington, D.C., 1968.

[20] M. K. Buckland, A. Hindle, A. G. Mackenzie and I. Woodburn: *Systems analysis of a university library.* (University of Lancaster Library Occasional Papers, 4). Lancaster, England, University of Lancaster Library, 1970. (ISBN 0-901699-01-2; ERIC report ED 044 153) p. 50.

[21] M. K. Buckland: *Book Availability and the Library User.* Chap. 7. Pergamon Press, New York (1975).

[22] *Liaison*, p. 49. Insert in: *Library Association Record* **73**(8) August 1971..

Other references relating to overlap and coverage

[23] R. O. Beauchamp, M. A. Daugherty, J. L. Garber and J. D. Myers: Comparative searching of computer data bases. *J. Chem. Docum.* 1973, **13**(1), 32–35.

[24] T. Besterman: The European union catalog project. *J. Docum.* 1958, **14**(2), 56–64.

[25] C. P. Bourne: *Overlapping coverage of Bibliography of Agriculture by* 15 *other secondary services.* Information General Corp., Palo Alto, Cal., 1969. (PB 185 069).

[26] J. L. Carmon and M. K. Park: User assessment of computer-based bibliographical retrieval services. *J. Chem. Docum.* 1973, **13**(1), 24–27.

[27] T. K. Devon, J. S. Buckley, E. D. Taylor and M. E. D. Koenig: Comparative evaluation of Ringdoe and CBAC. *J. Chem. Docum.* 1973, **13**(1), 30–32.

[28] C. K. Elliott: Abstracting and indexing services in psychology: a comparison of "Psychological abstracts" and "Bulletin signaletique". *Library Association Record* 1969, **71**(9), 277–8.

[29] C. M. Flanagan: Coordination—A detailed review of the relationships among the publications and services of BIOSIS, CAS, and Ei. *J. Chem. Docum.* 1973, **13**(2), 57–59.

[30] E. Garfield: *Article-by-article coverage of selected abstracting services.* Institute for Scientific Information, Philadelphia, Pa., 1964.

[31] A. Gilchrist and A. Presanis: Library and Information Science Abstracts, the first two years. *Aslib Proceedings* 1971, **23**(5), 251-256.

[32] D. J. Goode, J. K. Penry and J. F. Caponio: Comparative analysis of Epilepsy Abstracts and a MEDLARS bibliography. *Bulletin of the Medical Library Association* 1970, **58**(1), 44–50.

[33] R. C. Greer and P. Atherton: *Study of Nuclear Science Abstracts and Physics Abstracts coverage of Physics journals.* (AIP/DRP 66–11). American Institute of Physics, 1966.

[34] S. Jérôme: Comparative study of the coverage of physics journals by two computerized data bases—SPIN (Searchable and Physics Information Notes) and CAC (Chemical Abstracts Condensates). *Inform. Stor. Retr.* 1973, **9**(8), 449–455.

[35] J. Martyn: Tests on abstracts journals: coverage overlap and indexing. *J. Docum.* 1967, **23**(1), 45–70.

[36] J. Martyn and M. Slater: Tests on abstracts journals. *J. Docum.* 1964, **20**(4), 212–35.

[37] R. R. Miller: A study of searching the eye literature. *Am. Docum.* 1968, **19**(3), 223–239.

[38] R. H. Orr: The metabolism of information in psychopharmacology. *Psychopharmacology Service Center Bulletin* July 1961, 4–6.

[39] R. M. Orr and E. M. Crouse: Secondary publication in cardiovascular, endocrine and psychopharmacologic research. *Am. Docum.* 1962, **13**(2), 197–203.

[40] D. R. Smith, R. O. Beauchamp, J. L. Garber and M. A. Daugherty: Computerized drug information services. *J. Chem. Docum.* 1972, **12**(1), 9–13.