Intellectual Access to Archives:

II. Report of an Experiment Comparing Provenance and Content Indexing Methods of Subject Retrieval

RICHARD H. LYTLE

1. Introduction

IN A PREVIOUS ARTICLE in the *American Archivist,* the Provenance and the Content Indexing methods of gaining subject access to archives were described.[1] The descriptions made precise distinctions between the methods, primarily to derive method characteristics that could be further explored in a retrieval experiment.

In this article, the experiment will be described, its results presented, and implications traced for archival subject retrieval systems. The article is organized to assist those who wish to be selective in reading data and details of the experimental design. In the remainder of this introductory section, experimental methodology will be discussed and the experiment itself briefly described. Major findings and implications of the study are presented in Section 2. Details on research design and conduct of the experiment are presented in Section 3. In Sections 4 and 5, detailed findings are presented.

1.1. *Why Information Retrieval Experiments?* The very idea of a retrieval experiment in an archives is so unusual—probably this is the first—that the notion itself requires discussion. Is such an experiment feasible, in the sense that results may be generalized beyond the immediate experiment and thus have practical value?

---

[1] Richard H. Lytle, "Intellectual Access to Archives: I. Provenance and Content Indexing Methods of Subject Retrieval," *American Archivist* 43 (Winter 1980): 64-75 (hereafter referred to as "Access I"). Both articles are based on my doctoral disseration; see Richard H. Lytle, "Subject Retrieval in Archives: A Comparison of the Provenance and Content Indexing Methods," Ph.D. dissertation, University of Maryland, 1979, hereafter referred to as "Subject Retrieval." The two methods were defined in the first article as follows: The Provenance or P Method is the traditional method of archival retrieval, based on principles of archives administration and reference practices of archivists. Subject retrieval in the P Method proceeds by linking subject queries with provenance information contained in administrative histories or biographies, thereby producing leads to files which are searched by using their internal structures. The second method, the Content Indexing or CI Method, derives from librarianship but has been applied extensively to manuscript collections, and, to a limited extent, to archives. Subject retrieval in the CI Method matches subject queries with terms from an index or catalog. See "Access I," page 64.

Information retrieval experiments in libraries and other information systems have, of course, been done. These experiments may be described as either *laboratory* or *real-world* experiments. In a laboratory experiment, attempts are made to control all variables so that outcomes may with some certainty be ascribed to the variables of interest; for example, in a laboratory experiment to compare P and CI methods, the respective access devices would be made equal in quality to facilitate comparison. Laboratory experiments have the advantage of greater protection against influences from unknown variables; but many laboratory experiments are so far removed from real information systems that what can be concluded has no practical value. A real-world experiment, of which the present is an example, is conducted in a real information system setting. But the real-world retrieval experiment has liabilities, common to social-science field research generally, including limited ability to manipulate independent variables, limited ability to unscramble interactive effects of independent variables, and, probably, effects of unknown variables. For example, one is unlikely to find equal P and CI access devices for the same collections—which in fact was the case in this experiment—and this limitation will render less certain whether effects observed are due to method characteristics or peculiarities of the experimental environment.

The experiment reported here suffers from a number of the limitations common to real-world experiments. It also suffers from some limitations imposed by resources, most notably a lack of staff for administering the experiment. From a research design viewpoint, the most severe restriction was the small number of questions, a restriction opening the experiment to possible distortion from unrepresentative questions. But the experiment does advance our knowledge of P and CI methods and, perhaps more important, it demonstrates that a sound research design is feasible for comparing archival subject retrieval systems.

1.2. *Summary of the Experiment.* The experiment to explore subject retrieval capabilities of Provenance and Content Indexing methods was carried out in 1978 at the Baltimore Region Institutional Studies Center (BRISC), a division of the University of Baltimore.[2] BRISC collects archives and papers relating to metropolitan Baltimore for urban studies research; most of its collections come from the Planning Department of the City of Baltimore and from Baltimore civic groups such as the Greater Baltimore Committee.

The general methodolgy of the experiment was to run the same questions using Provenance and Content Indexing methods and compare the results. Fifteen questions were run, twice each on the two methods for a total of 60 runs. Ten questions were selected from BRISC's record of past questions (dead questions), and five were received from current BRISC users (live questions). Some 897 folders were retrieved in response to the fifteen test questions. Fifteen collections comprising approximately 380 cubic feet were used in the experiment, selected from total BRISC holdings because they were accessible by both P and CI meth-

[2] I owe special thanks to BRISC for permitting me to do the experiment, and for substantial staff support in its execution. Especially helpful were Director W. Theodore Dürr, and Associate Director Adele Newburger. Others who assisted in the experiment were: D. Randall Bierne, Richard Cox, Richard Szary, Jerry Watkins, and Karen Womble. Following this article are Dr. Dürr's comments on the experiment.

ods. They had been indexed by BRISC[3] and at least rudimentary P Method finding aids existed. Four searchers were used in the experiment, including one present and one past BRISC staff member, one manuscript curator, and one federal archivist. Five users presented live questions and rated those folders. A judge, an associate professor of sociology at the University of Baltimore, rated folders retrieved for both live and dead questions.

The reader should bear in mind that the purpose of the experiment was a study of P and CI methods, not an evaluation of BRISC. Had BRISC evaluation been the purpose, some key aspects of the research design would have been changed, which in turn would have altered the results.

2. Major Findings and Implications of the Study.

2.1 *Major Findings.* For those archivists who have suspected that they are retrieving only a small fraction of relevant documents for users, the results of this study are confirming evidence. The most salient finding of the study was the poor retrieval performance of both methods. The conclusion that both methods performed poorly was drawn by measuring overlap, the degree to which both methods found the same relevant folders. Since overlap was low, it is improbable that the four searchers found most of the relevant folders in the collections; in other words it is likely that many relevant folders were not retrieved. Another way of stating this finding is that neither method is very consistent or reliable, as measured by overlap. Although it is possible that this finding is peculiar to BRISC,[4] that seems unlikely. Probably this result is typical of retrieval from archives.

The relative retrieval performance of Provenance and Content Indexing methods, averaged across the experiment, was approximately equal, although the CI Method exhibited more variance (had more high and low scores) than did the P Method. This was a surprising finding, considering the inadequacy of P Method finding aids and the apparent high quality of the CI Method indexes at BRISC. The P Method may have greater strengths than even its advocates have imagined.

The CI Method average retrieval performance for the entire experiment was affected considerably by two or three low scores, including one very low score. The reason for the extremely low score was absence of terms representing the user's concept in the system vocabulary; i.e., there were no entries in the index for the subject the user was seeking. When the desired concept is not represented in the system vocabulary, one is better off with the *ad hoc* procedures of the P Method. This result is hardly surprising in principle, but it constitutes a warning for those designing archival subject access systems, where demands on the system

---

[3] W. T. Dürr, ed., and P. M. Rosenberg, comp., *The Urban Information Thesaurus, a Vocabulary for Social Documentation* (Westport, Connecticut: Greenwood Press, 1977). The term *collection* is used in this article and in "Access I" to indicate the intellectual control unit, close in meaning to Schellenberg's term, *record unit.* For the P Method at BRISC, each accession/series was considered a collection for these purposes, although BRISC refers to them as series. None of the definitions established in the glossary of archives terminology precisely fits this concept; see *American Archivist* 37 (July 1974): 415–33.

[4] This qualification could be applied to all of the findings in the experiment and was discussed in Section 1. Note also that the findings do not necessarily reflect what BRISC would have retrieved; they use both methods in reference service.

are poorly defined.[5] On the other hand, if the system vocabulary can anticipate user demands, the CI Method has considerable potential as indicated by its high scores in this experiment.

The experiment revealed a significant relationship between searcher and method performance. The most important factor was experience on method. Retrieval performance scores appeared to be primarily a function of the experience of the searcher with the method used. A searcher experienced in the P Method achieved good results with it, as did a searcher experienced in CI using that method. System designers must consider the searcher, either by providing training or by designing the system to accommodate searcher weaknesses.

Because of limited resources and pecularities of the experimental environment, several variables of interest were not studied. The most important of these was the number of files to be accessed. A large number of files strains the P Method approach, but the CI Method alternative becomes quite expensive when applied to large accumulations of records. Testing the "number of files" variable would require a much larger number of files accessible by both methods than exists at BRISC, and probably no repository exists which is suitable for such a test.[6]

2.2. *Implications of the Study.* The most immediate and compelling implication of this study is that archivists should evaluate their information retrieval systems. The methodology developed for comparing Provenance and Content Indexing could, with modifications, be used to evaluate and compare any two, or more, access systems. Performance evaluation should be a component of most subject-retrieval design projects, and perhaps should be required by funding agencies in most instances. Results should be reported in the literature.

Evaluation of existing information retrieval systems has the liability that basic assumptions of these systems may go unexamined. The systems analysis undertaken in "Access I" suggests a wider approach that is less dependent on the existing system. The system must be defined, including its users and what they demand of the system.[7] All of these considerations should be taken up within the fullest possible cost benefit context. Subject access systems should be constructed with a reasonable expectation that they will serve a user need, and that the cost of serving that need is in proportion to the importance of the need and the resources of the system. Costs have not been discussed here, but obviously they are quite important.

The study indicates that research is needed on how the Provenance Method retrieves from archives. Specifically, research is needed to define (a) how archi-

[5] "Access I," pages 69–70.

[6] The "number of files" problem can be further elaborated. The P Method leads to increasing difficulties with increasing numbers of files, and the problem becomes acute in large repositories such as state archives or the National Archives. P Method difficulties are of two types: (1) When the number of files is very large, the problem of selecting files for searching becomes critical; and (2) When the P Method identifies many files as of probable interest, the problem of searching many files may be overwhelming. The CI Method offers theoretical relief in this situation, but it is precisely in such large bodies of records that the cost of applying the CI Method becomes exhorbitant.

[7] See "Access I," pages 64–66.

vists using the P Method proceed from a subject question to the files to be searched; (b) how effective the P Method is in aiding searches within the files; and (c) whether these P Method approaches can be systematized to the extent that they can be computerized. Perhaps an improved version of the P Method would be the most cost-effective retrieval device for the archives system.

Application of the CI Method to the archives system should focus on definition of user needs to overcome the severe problem which arises when the user's concepts are not included in the system vocabulary.

Implicit in many of the preceding observations is the need for more research in archives administration. Although the profession has made great strides in teaching archives administration, archivists do very little research comparable to library science research conducted in some library schools. The archival profession should develop a tradition of theoretical and empirical research, certainly in the areas of information systems and appraisal.

3. Details Concerning Design and Conduct of the Experiment

3.1 *Factorial Design.* The major design of the experiment consisted of sixty cells, each consisting of one run of one question; there were fifteen questions, each run four times, twice on each method. Ten of the questions were selected from past BRISC questions and five were received from current BRISC users during the experiment.

A major concern of the design was to distinguish between BRISC and non-BRISC searchers, to guard the experiment from possible bias; the research design permitted comparison of BRISC versus non-BRISC on P versus CI methods, a comparison which might not have been possible had the research design not been so constructed.

Within the limited resources available, there was a choice to run more questions twice, once by each method, or to run fewer questions four times, twice by each method. Four runs, two on each method, was the option selected to increase reliability—to reduce variability introduced by searcher characteristics; of course, that solution opened the experiment to greater variation introduced by individual questions. Each searcher ran the fifteen questions in sequence, alternating methods; one reason for this was to control for learning effects, so that as searchers unacquainted with BRISC gained experience, that increasing experience was evenly distributed across methods.

3.2. *Method Access Devices.* The Provenance Method access devices consisted of collection-level descriptions and folder lists—of rather low quality as a result of the BRISC policy to allocate more resources to indexing than to improvement of P Method finding aids. The Content Indexing Method devices were based on a machine-readable data base created by indexers using the *Urban Information Thesaurus.* Because of software limitations, most searches were done from computer printed indexes and, in a few cases, on-line searches.[8] Clearly, P Method access devices at BRISC rated substantially below CI Method devices; but the CI Method suffered too from a restriction against Boolean searching. The comparison cannot, however, be quantified to weight performance measures in the experiment.

[8] Dürr and Rosenberg, *Urban Information Thesaurus.* Although CI Method searchers could use the system on-line, budget limitations had prevented BRISC from developing software to make Boolean searches. Thus, except in a few cases (for example, some geographical qualifications of subject queries) on-line searches provided no greater searching power than did the index.

3.3. *Questions.* A typology of subject queries was established by examination of the past record of BRISC search questions. The typology and question examples are given in Table 1.

Table 1
Question Typology and
Question Examples

1. *Specific documents*
   I want to see Planning Commission agendas.
2. *Geographical specification*
   What do you have on the planning or construction of the Bridge/Tunnel at Ft. McHenry?
3. *Strong institutional specification*
   What materials do you have concerning the Model Urban Neighborhood Development projects, in addition to MUND records themselves?
4. *Weak institutional specification*
   How was the decision made to sell Friendship Airport to the State of Maryland?
5. *No geographical or institutional specification*
   What was women's involvement in neighborhood organizations?

The typology of questions was related to a hypothesis about the performance of P and CI methods: that the P Method will perform its best in comparison with the CI Method for questions containing strong institutional references, while the CI Method will perform better than the P Method for questions without institutional references. The typology does have its weak points. For example, it is difficult to distinguish between weak and strong institutional specification; and implicit institutional specification may be strong in a question which otherwise is without such specification.

3.4. *Searchers, Judge, and Users.* Four searchers were used in the experiment. Searcher 1 had an M.A. in history and primary experience as a manuscript curator in a historical society. Searcher 2 had a B.A. and previously worked at BRISC where she cataloged several of the collections used in the experiment and performed reference service. Searcher 3 had an M.L.S. and had worked at BRISC since its inception; she was thoroughly knowledgeable about BRISC and its systems and about the collections used in the experiment. Searcher 4 was an experienced mid-level archivist in the federal service; holding an M.A. in history, he had extensive experience in searching archives by the P Method, very little contact with CI Method systems, and little background in Maryland and Baltimore history.

The judge, who had a thorough knowledge of Baltimore history, rated folders retrieved for all questions. The five users who presented live questions and rated those folders were two University of Baltimore undergraduates, a foundation official, a City of Baltimore official, and a sociology professor at the University of Baltimore. Judge and users rated each archives folder on a scale of one to four, with "1" indicating highly relevant and "4" indicating not relevant.

3.5. *Independent Variables.* Table 2 presents a list of independent variables identified for study in "Access I." These variables are characteristics of the archives system pertinent for study of subject retrieval by P and CI methods. As noted in the table, not all of the variables were in fact the subject of study in the experiment, due to limited resources and peculiarities of the experimental environment. The variables will not be discussed here, but the reader may find them explained in "Access I."[9]

[9] See "Access I"; much of that article identifies and describes the independent variables. Although the form of presentation there varies considerably from the present table, the reader can find independent variables discussed, sometimes in detail. Greater detail yet can be found in "Subject Retrieval," especially pp. 45–52.

Table 2
Independent Variables Identified for Study

Variables Pertaining to the Access Mode
- Provenance versus Content Indexing Methods*

Variables Pertaining to the Collection
- Order versus Disorder in the Collection
- Informational Context of the Collection
- Size of the Collection
- Number of Collections to be Accessed*
- Permanent Retrospective Character of Collections
- Use Rate of Collections

Variables Pertaining to Collection Preparation
- Quality of Finding Aids
- Quality of Indexing

Variables Pertaining to the Questions
- Type of Question*

Variables Pertaining to the User
- Type of User*
- Place on the Browsing Continuum

Variables Pertaining to the Intermediaries
- BRISC versus Non-BRISC Searchers*
- Experience versus Inexperience on Method*

Precision

**Variables represented in the experiment*

3.6. *Dependent Variables.* Dependent variables, listed in Table 3, are retrieval system performance measures. The dependent variables have numeric values, usually from a low of 0 to a high of 1. Some of the variables are mathematically related to each other.[10] Also, all of the dependent variables are *comparative,* in the sense that they measure relative performance of P and CI methods; they do not state method performance measured against some third or absolute standard. The individual performance measures are explained as the data are presented.

Table 3
Dependent Variables Used in the Experiment

Recall-Based Variables
- Comparative Recall
- Overlap Within Methods
- Overlap Between Methods
- Incremental Advantage Measure
- Efficiency

Working Time

In order to compute performance measures such as comparative recall, the 1–4 rating scale was collapsed into a dichotomy of relevant (1 or 2) and not-relevant (3 or 4); i.e., all folders rated 1 or 2 were counted as relevant and all folders rated 3 or 4 were counted as not relevant.

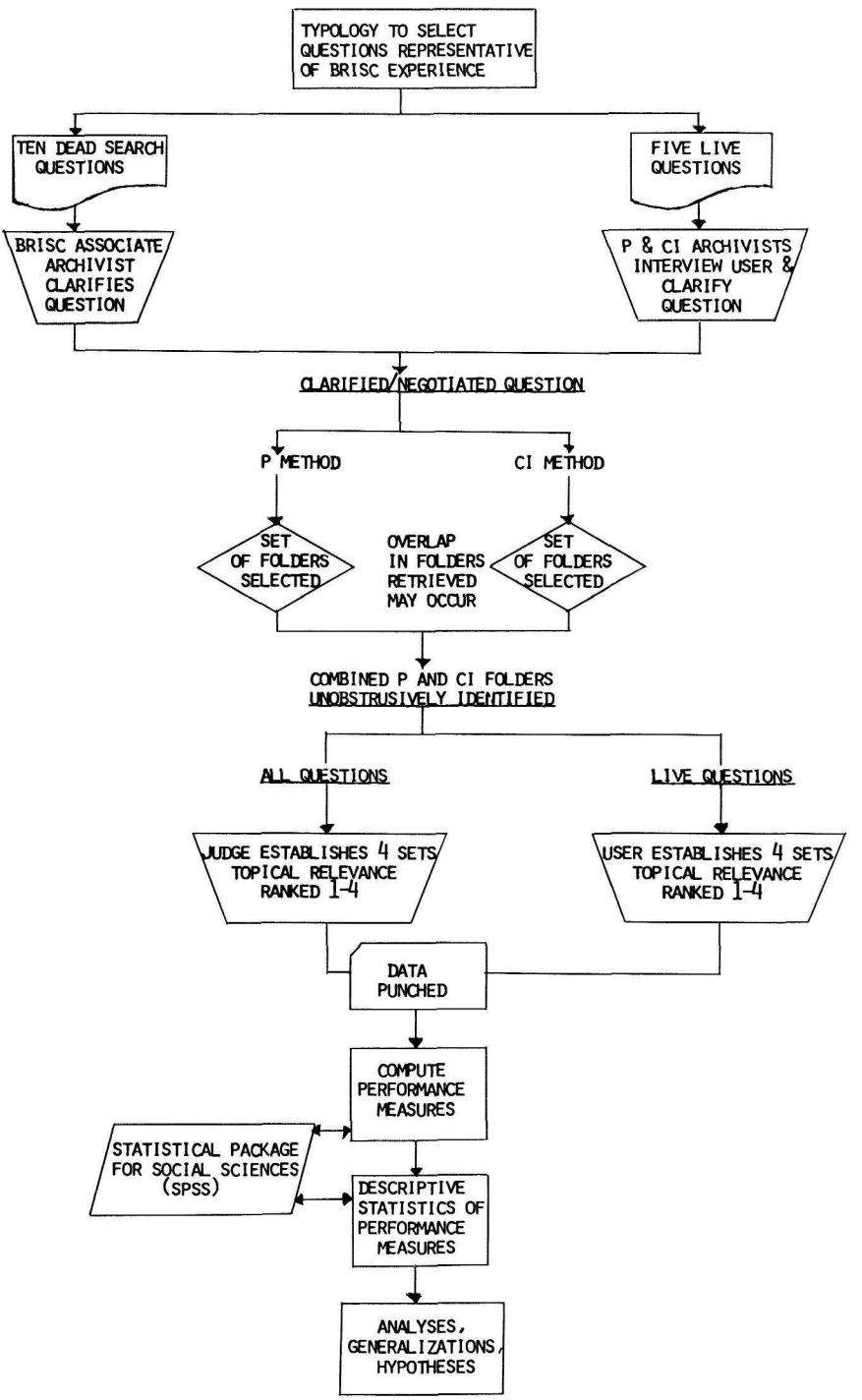
3.7. *Conduct of the Experiment.* Protocols established how searches and ratings were to be accomplished; the flow of the experiment is summarized in Table 4. The experiment was designed to separate P and CI methods as completely as possible in retrieval, combine the results (retrieved folders), and present the combined set for rating without identifying which method retrieved individual folders. Performance measures were then calculated based on the ratings and which method did in fact retrieve relevant folders.

Access devices were segregated by method. Each searcher retrieved folders for one question using only P or CI devices,

[10] The mathematical relationship problem is discussed in "Subject Retrieval," pages 65–66, and I plan to publish details elsewhere. For information on calculation of similar performance measures, see Tefko Seracevic, et al., *An Inquiry into Testing of Information Retrieval Systems, Pt. II, Analysis of Results* (Cleveland: Case Western Reserve Center for Documentation Research, 1968), pp. 8–10. Some readers will note that a customary performance measure, precision, is not included here. Precision indicates what percentage of folders retrieved are relevant, or conversely, the proportion of irrelevant folders a user must search to locate relevant folders. In this experiment, precision could not be used to measure *method* performance because the searcher had screened folders as he proceeded; especially on P, many more folders were examined than were retrieved. Working time was used as a partial substitute for precision.

Table 4

CONDUCT OF THE EXPERIMENT



finding aids for the P Method and computer-generated indexes for the CI Method. When using the P Method, searchers could browse folder headings in boxes, but not examine index terms recorded on the folders. When using the CI Method, searchers could not browse the file folder headings; but they could refer to index terms applied to folders already located, and go back to the index with additional terms for searching.

Establishment of a methodology which segregated methods served well the purpose of studying those methods, but it should be noted that by the same token it reduced the power of the experiment as an evaluation of BRISC.

3.8. *Reliability of the Experiment.* The judge's ratings provided most of the data for the experiment, and for that reason it was very important to establish the validity of his judgments; this was a major aspect of assessment of the experimental results. Only a brief overview will be included here.[11]

The judge was seen to be consistent in his ratings of dead questions (1–10) and live questions (11–15), and therefore his overall reliability was established by comparing his ratings to those of users (live questions only). Results indicated that (1) if differences of one point were ignored, there was 76 percent agreement between judge and user; and (2) the judge tended to be more lenient, tended to rate the folders more relevant, than the user did. Questions contributed unequally to judge-user agreement, with imprecise questions and questions asked by undergraduates contributing more than their share to the disagreement.

Although standards for comparing judge-user agreement results with other experiments do not exist, it was concluded that judge-user agreement was sufficient to ensure reliability of the experiment.

The data indicated no bias from BRISC or non-BRISC searchers.

4. *Detailed Findings: Quantitative Analysis*

4.1. *Poor Performance on Both Methods.*

4.1.1. *Summary.* The most salient result of the experiment is the finding that neither method is very consistent or reliable. The performance measure most indicative of this finding is overlap, which was quite low both between runs of the same method and between runs of different methods. Since low overlap indicates that each of the four runs found a mostly distinct set of relevant folders, probably many more relevant folders remained unretrieved. The combined performance of the two methods was poor.

4.1.2. *Data.* Overlap measures vary from 0, indicating no folders found in common, to 1, indicating complete overlap or all folders found in common.

Overlap within methods indicates the extent to which two runs by the same method found the same relevant folders. This measure is calculated by dividing the number of relevant folders retrieved in both runs of a method by the number of relevant folders retrieved in either run of a method. (If there is any overlap, the denominator will be larger than the numerator.) Data from the experiment indicated that *overlap within P was .36* (standard deviation .31) and *overlap within CI was .41* (standard deviation .37). Both methods were low on overlap within methods, and the difference of .05 was not regarded as important.

Overlap between methods is calculated much like overlap within methods, except that numerator and denominator sets are collapsed between runs of each method before the overlap calculation is made (i.e., a given folder is counted a maximum of one time, even if it is retrieved by both runs of one method). Overlap between methods, then, is calculated by dividing the number of folders retrieved in both methods by the number of folders retrieved in

[11] See "Subject Retrieval," pages 95–109, for much more detail concerning judge-user agreement and reliability of the experiment.

either method. (If there is any overlap, the denominator will be larger than the numerator.) *Data from the experiment show that overlap between methods was .36* (standard deviation .32). Overlap between methods was low.

4.2. *Equal Overall Retrieval Performance as Measured by Comparative Recall*

4.2.1. *Summary.* Summarized across the entire experiment, retrieval performance was approximately equal for Provenance and Content Indexing methods. The performance measure used was comparative recall, which states the comparative retrieval power of P and CI Methods; comparative recall values for each run of each method on each question were averaged to give values for each method across the experiment. These values were approximately the same.

This finding was quite surprising, given the emphasis of BRISC on CI Method access devices and the relatively poor quality of P Method finding aids. It was noted, however, that while the averages were close, variation as measured by standard deviation was quite different. The CI Method exhibited considerably greater variation; that average contained many high and low values, while values for the P Method were much closer to the mean. These data are further explored below, especially in Section 4.7.

4.2.2. *Data.* Under ideal circumstances, recall would measure the percentage of all relevant folders in the entire collection retrieved by the method of interest; the problem is that the number in the entire collection of folders relevant to a question is not known. Comparative recall uses as the denominator the number of relevant folders retrieved by all four runs of a question, counting a folder only once no matter how many times retrieved; this number clearly is not a good estimator of the actual number of relevant folders in the collections, but it does permit construction of a comparison measure. Comparative recall, then, is the number of relevant folders retrieved by one run of one question, divided by the denominator just described.

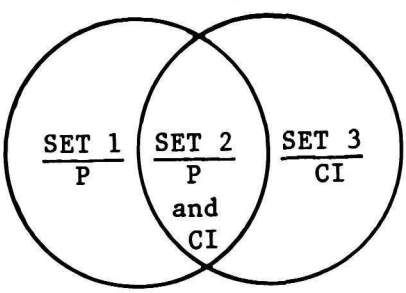
Comparative recall was approximately equal for P and CI Methods; P was .48 and CI was .50. The standard deviation for comparative recall did vary by method, however; standard deviation for P was .24 and for CI, .33. This indicates that the CI average consisted of more high and low scores, while the P Method scores were closer to the .5 mean.

4.3. *Similar Performance on the Incremental Advantage Measure.*

4.3.1. *Summary.* Given retrieval by one method, what advantage is gained by adding folders retrieved by the other method? A performance measure, called the incremental advantage measure, was devised to answer this question. Values of the incremental advantage measure vary from 0 (for example, no advantage of adding P retrieval to the CI retrieved set of folders) to numbers greater than one (for example, "considerable" advantage of adding P Method retrieval to the CI retrieved set of folders).

Although the data on this measure are somewhat difficult to interpret, no large differences in method performance were evident.

4.3.2. *Data.* The incremental advantage measure, although conceptually simple, requires more explanation than previous performance measures. The Venn diagram indicates the folder sets involved in the incremental advantage measure.



Sets 1 and 3 include those folders retrieved by one method but not by both; set 2 includes those folders retrieved by both methods. Set 1 is the incremental advantage of adding P Method retrieval, given CI; and set 3 is the incremental advantage of adding CI Method retrieval, given P.

The incremental advantage measures are calculated as follows:

(1) Incremental Advantage of P, Given CI Retrieval

$$\frac{\text{set 1}}{\text{set 2} + \text{set 3}}$$

(3) Incremental Advantage of CI, Given P Retrieval

$$\frac{\text{set 3}}{\text{set 1} + \text{set 2}}$$

Previously reported performance measures were calculated by using a single relevant/not-relevant dichotomy derived from the 1 to 4 rating scale: relevant (1 or 2) and not-relevant (3 or 4). This was a simplification of actual performance measure calculations in the experiment. Actually, every performance measure was calculated at three relevance levels, as follows: Level 1 (relevant=1, not-relevant=2, 3, 4); Level 2 (relevant=1, 2; not-relevant=3, 4); Level 3 (relevant=1, 2, 3; not-relevant=4). All three levels are introduced here, because only for the incremental advantage measure were any important differences in method performance noted at the different relevant levels.

Data for the incremental advantage measures are shown in Table 5.

Table 5
Incremental Advantage Measures
Average (Standard Deviation)

	Level 3	*Level 2*	*Level 1*
Incremental Advantage of P Given CI	.71 (.84)	.79 (.97)	1.16 (1.78)
Incremental Advantage of CI Given P	.70 (.31)	.56 (.47)	.58 (.78)

Note: A value of 0 indicates no incremental gain; a value larger than 0 indicates some incremental gain.

The data indicate approximately equal performance at Level 3, but increasing incremental advantage of P at Levels 2 and 1.

Raw data showing the *number of questions* for which retrieval was greater for each method are shown in Table 6. These data indicate the same general trend as the incremental advantage measure: as relevance requirements are made more stringent, the incremental advantage of P increases.

4.4. *Effect of Detailed Analysis on CI-Retrieved Folders.* The experimental findings noted in the preceding paragraphs—comparative recall, overlap measures, and incremental advantage measures—indicate approximately equal performance of the methods. This result was unexpected; it was anticipated that the CI Method would outperform the P method. Because of the investigator's surprise at the outcome, a tentative hypothesis was tested, as follows: CI Method indexers may perceive relevance in a folder which the user or judge, especially on hurried analysis, may overlook. The indexer may have greater subject knowledge than the user and if trained properly may be more sensitive to relevant aspects of folders.

To test this hypothesis, a random sample of folders retrieved by P versus CI methods—selected by the computer—was evaluated for changes of the judge's opinion. The judge was specifically instructed to take his time in evaluating this small group of folders. The judge did raise his assessment more often on the CI-selected folders than on the P-selected folders. Since the number of folders was small, and since most changes were by one point only, conclusions are not possible. But interest in the hypothesis is strengthened: the CI Method, more than the P Method, may retrieve folders which the user will consider relevant only upon careful analysis.

4.5 *Slightly Better Working Time Performance by CI Method.* A final performance measure calculated for each method across the ex-

Table 6
Incremental Advantage Measure
Number of Questions

Level	Tie	Gain by P Given CI	Gain by CI Given P
1	3	7	4
2	2	6	6
3	1	5	8

periment was working time, a measure of the amount of time spent in searches by each method. The CI Method consumed somewhat less working time than did the P Method—across the experiment, an average of some 60 minutes per question contrasted with 69 minutes for the P Method.[12]

4.6. *Comparative Recall a Function of Searcher Experience.*

4.6.1. *Summary.* No archivist will be surprised to learn that searchers were an important factor in method performance. Comparative recall appears to be primarily a function of the experience of the searcher with the method used, rather than a function of the method. A searcher experienced in the P Method will achieve good results with it, as will a searcher experienced in CI using that method. Moreover, a searcher needs less time to achieve a given result when he uses a method in which he has experience.

4.6.2. *Data.* Comparative recall calculations have been explained above. Comparative recall was averaged for each searcher on each method to give two values for each searcher: recall on P and recall on CI. Recall was also adjusted by working time to raise recall values for brief searches and to lower recall values for lengthy searches. This measure is called *efficiency*.[13] Table 7, below, gives data for recall and efficiency averaged for searchers experienced on method versus searchers inexperienced on method. These data are the basis for the conclusion that interaction of searcher and method was a very important aspect of the experimental results.

Data for each searcher, given in Table 8, below, are also of interest. Searcher 1 (non-BRISC), who had a manuscript reference background, was inexperienced on both methods as represented in this experiment. Also, he spent less time on searches, which resulted in low recall figures, especially on CI, but which also resulted in considerably

Table 7
Recall and Efficiency
Experience versus Inexperience on Method

	Searcher Experienced on Method		Searcher Inexperienced on Method	
	Recall	*Efficiency*	*Recall*	*Efficiency*
Provenance Method	.56	.68	.45	.46
Content Indexing Method	.56	.75	.47	.51

Note: values of efficiency may vary from 0 to numbers greater than one.

[12] One extremely long run caused distortions in the averages, and thus the longest run on each method was eliminated from the calculation.

[13] Efficiency was calculated by dividing recall by working time, and multiplying that result by a measure of question difficulty. The measure of question difficulty was a specially calculated average of the working times of the four runs of a given question.

Table 8
Recall and Efficiency by Searcher

Searcher	Provenance Method Recall (std.dev.)	Provenance Method Efficiency (std.dev.)	Content Indexing Method Recall (std.dev.)	Content Indexing Method Efficiency (std.dev.)
1. Non-BRISC Naive on P and CI	.42 (.19)	.59 (.46)	.29 (.29)	.45 (.56)
2. BRISC Knowledgeable on CI	.48 (.32)	.51 (.28)	.49 (.32)	.98 (.85)
3. BRISC Knowledgeable on CI	.45 (.08)	.27 (.10)	.63 (.29)	.51 (.24)
4. Non-BRISC Knowledgeable on P	.56 (.29)	.68 (.49)	.64 (.34)	.57 (.33)

Note: values of the efficiency measure may vary from 0 to numbers greater than one.

upgraded values on efficiency. Searcher 2 (BRISC) did reasonably well on all searches and, considering lower search times, had the best performance on the CI efficiency measure. (Note that efficiency can have values greater than 1; this accounts for the high average and standard deviation of Searcher 2.) Searcher 3, presently a BRISC staff member, was downgraded for longer searches on the efficiency measure, but, as might be expected, performed best on CI as measured by recall. Searcher 4 (non-BRISC) was experienced on P and performed best on that method, although he also performed well on CI.

4.7 *Importance of Questions in Experimental Results.* There was a great deal of variance in retrieval performance by question, and there was greater variance across questions on the CI Method than on the P Method. Quantitative analysis failed, however, to provide a satisfactory explanation for that variation.

A typology of BRISC search questions, presented in Table 1 above, was developed to test the following hypothesis: the P Method will perform its best in comparison with the CI Method for questions containing strong institutional references, while the CI Method will perform better than the P Method for questions without institutional specification. Experimental results failed to reveal notable variation explained by this typology, and thus the initial hypothesis was not supported. Moreover, no classification of questions attempted produced marked differences in average recall figures.

Ranking of questions according to performance on comparative recall did produce results; there was a tendency for methods to rank high on the same questions but low on different questions.

Generally speaking, however, quantitative analysis did not provide acceptable explanations for the fact that P and CI methods performed poorly on different questions.

4.8. *Importance of Collections.* The CI Method accessed more collections, on the average, than did the P Method. Data for all folders retrieved (i.e., including those judged not relevant) indicate that the P Method accessed an average of 4.4 collections per question (range: 1–7 collections), while the CI Method accessed an average of 5.6 collections per question (range: 1–10 collections). Clearly, the CI Method, on the whole, retrieved from more collections than did the P Method. However, on those questions for which the CI Method accessed notably more collections than the P Method, CI retrieval was generally lower than P Method retrieval. This suggests the hypothesis that CI selects collections for searching better than P, but that the P Method does better searching within a collection. Limitations on study of the collection variable in this experiment, however, render this hypothesis quite tentative.

5. *Detailed Findings: Failure Analysis*

5.1. *Introduction.* The purpose of failure analysis is to provide explanations for quantitative performance data. Failure analysis proceeds beyond quantitative analysis to explore the *whys* of system success or failure. For example, folders retrieved by CI but not by P are examined, along with finding aids, to explain the P Method retrieval failure. The methodology of the failure analysis in this experiment was quite simple. *Questions* were selected, based on recall figures, as follows: for each method, the three best, three worst, and three most different. Because several questions occurred in more than one category, this procedure resulted in ten questions in all. The failure analysis was carried out on random samples from three sets of folders retrieved in response to the above questions. The three folder sets were as follows: selected by P and not CI; selected by CI and not P; selected by P and CI.

5.2. *Typology of System Failure and Examples of Failure Analysis Findings.* Although the general causes of information retrieval failure are known from other studies—for example, those of F. W. Lancaster concerning MEDLARS[14]—the specific characteristics are not known of Provenance and Content Indexing Method archives system failure. Thus, a major objective of the present analysis was to develop a typology of P and CI Method characteristics for future failure analysis. This typology is presented below in Table 9.

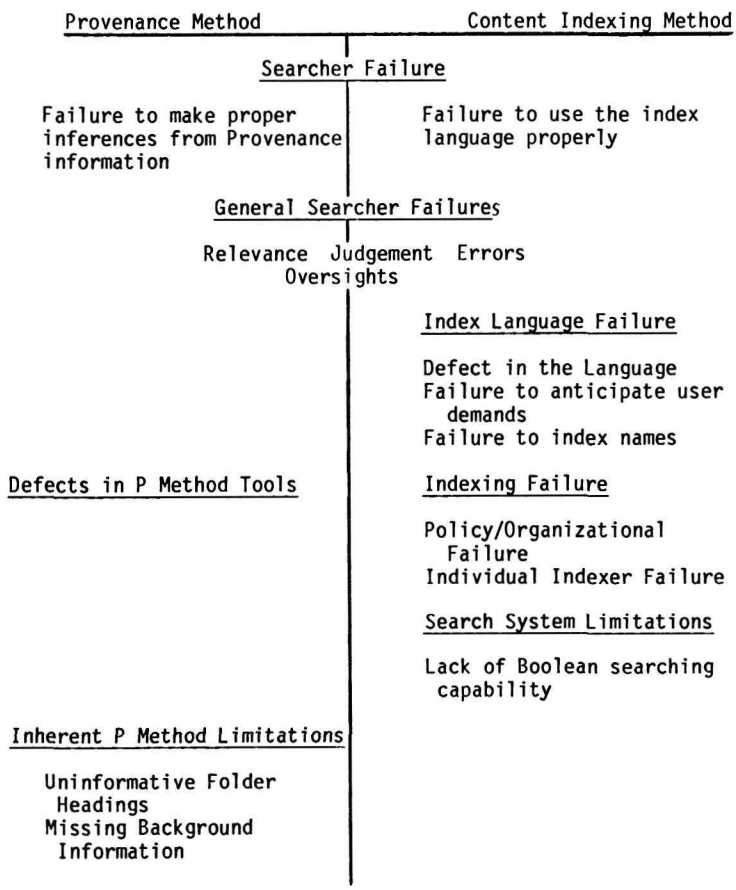
Searcher failure was a pervasive cause of retrieval failure in this experiment. On the CI Method, some searchers failed to use the index language properly, either missing relevant concepts entirely or selecting the wrong level of generality in the hierarchy. For example, on a question requesting folders pertaining to planning for and citizen reaction to the Ft. McHenry Bridge/Tunnel, one searcher used East-West Expressway and Tunnels/Bridges, retrieving thereby a great number of folders on actual construction and detailed plans for the construction, but little about citizen concern over the plan; the other searcher used descriptors and qualifiers such as Planning Process, Citizen Participation, and Social Aspect, which did retrieve more pertinent folders. Search formulations on the less precise questions were widely variant, a major factor in low overlap within the CI Method. On the P Method, there were several examples of failure to make proper inferences from P Method information. For example, on a question concerning the ABCD adoption project in Baltimore, the pertinent series description should have led the searcher to examine boxes which he entirely missed. Failure to make proper inferences from provenance information, and consequent failure to locate files for searching, is a classic example of subject searching failure in the P Method. Searchers on both methods committed screening errors: folders which were relevant were viewed but not selected.

A few examples of indexing failure were discovered, where a folder retrieved by the P Method but not by the CI Method was not indexed by important terms from the BRISC thesaurus. Moreover, some failures on CI clearly were due at least in part to budget-imposed searching system limitations at BRISC; searching on-line using combinations of terms was largely impossible. P Method failure due to defective finding aids was prevalent throughout the experiment; this problem became quite apparent when investigating why P failed to retrieve a folder found by CI. In many cases folder headings were nonexistent or were not entered in folder lists.

Failures due to intrinsic P Method limitations are more interesting than those due to defects. Uninformative folder headings were the most prevalent of these problems. An example is a folder labelled "Minutes of Meetings" in the Greater Baltimore Committee records; Content Indexing did indicate relevance to a question on the Model Urban Neighborhood Development project which the P Method would be unlikely to find except by searching the entire

[14] F. W. Lancaster, "Aftermath of an Evaluation," *Journal of Documentation* 27 (1971): 1–10.

Table 9
Typology of System Failure Found in the Experiment
Provenance versus Content Indexing Methods



file. (It is possible that detailed provenance information, such as dates of meetings during which the project was discussed, might indicate where to look in the minutes). There were other instances where the P Method searcher was required to have more information than the CI Method searcher; the P Method searcher failed, where presumably he would not have in the CI method where the indexer had in effect supplied background information by applying a descriptor. For example, on a question concerning the Mount Vernon Urban Renewal Area, the CI Method searcher retrieved folders relating to the Walters Art Gallery while the P Method searcher did not; the P Method searcher did not make the connection between the Walters and the Mt. Vernon area of Baltimore.

The Content Indexing Method is at a severe disadvantage when the critical concept of the subject request is missing from the system vocabulary. A notable example of that problem in the experiment was the following question:

> Please find formal or informal reports of the City Planning Commission, 1965-1972, on resource allocation. Interested in information on aesthetic or "humanistic" factors in resource allocation. (Humanistic loosely defined to exclude narrowly technical considerations).

That both methods would have severe difficulties with this question could be expected, but the CI Method performed much worse on it than did the P Method, even if one considers only the retrieval efforts of the BRISC searcher on CI. BRISC staff confirmed that the index language provided only meager approaches to the important qualifying concept. Given the absence of terms approximating "humanistic" in the index language, the browsing characteristic of the P Method provided a better retrieval result.

RICHARD H. LYTLE is the Archivist of the Smithsonian Institution.

COMMENT

WHEN DICK LYTLE SUGGESTED to Adele Newburger and me that he would like to examine the BRISC retrieval system in use for control of our archives, he stated that several other archives, when similarly approached, had discouraged him. As we were approaching the end of an early phase of our system and were hard at work on construction of an improved system (based on six years of experience), we welcomed his timely appraisal. At the same time we all realized that an "open door policy" was a gamble—we might not look very good—and also inopportune in the sense that our original system (the one he would use) did not make possible the sophisticated searches which can be performed. Limitations imposed by the then-current computer programs did not provide for full Boolean on-line post coordinate searches using the terms of our thesaurus. Dick, very charitably, suggested that we were courageous. We felt that he was equally, if not more, courageous as he intended to devise a model for examination of a controversial subject in a field (archives) where allegiance to the *status quo* had run headlong into a dynamic challenge posed by automation (the computer).

The goals of BRISC are similar to those of any other archives: to preserve and establish over collections from various organizations intellectual control that will respect the integrity of each separate organization (and the finding aids it may have devised) and at the same time provide users with ready access to the information desired from as many collections as possible. We try to do this while keeping in mind the rules of parsimony and the goals of precision. Therefore, we were pleased when Lytle said that he wanted to measure factors like accuracy, speed, relevance, and recall.

As expected, Dick ran into some methodological problems of his own. The size of his sample is the single greatest methodological concern; it will not be addressed here because my comments are about use of the data he obtained, but his foremost problem cannot be overlooked.

In Section 4.1 Lytle deals with overlap and notes the low scores. The problem is highly aggravated by the fact that at the time of his research the system would not perform Boolean searches. Lytle comments in footnote 7 that this limitation "reduced the power of the CI method." From the point of view of information systems management this limitation all but *destroyed* its power. Furthermore, the methodology used to distinguish between P and CI methods strengthened the experiment for Lytle's purpose. In fact, results of Lytle's search do indicate what will be found by the P and CI methods. Use of this controlled method was necessary for the experiment but does not correspond to many provider/user outcomes at BRISC. Thus it must be emphasized that BRISC was the location of the experiment and not a party, *per se*, to either approach.

One observation by Lytle is that the CI method required an average of less search time (9 minutes less per search). It suggests greater efficiency for users of archives. Part of BRISC's ARCHON II has just been tested in a complex oral history/theatre program and the access it provided to researchers, writers, and actors could have been provided in no other way.

Madeline M. Henderson, writing in the October issue of the ASIS *Bulletin,* suggests that the 1980s will be the era concerned about the high cost of human resources as the most expensive part of many human service systems. Use of the computer has become so pervasive throughout industry that people seldom question the combination of service and savings provided. The success of OCLC and similar systems suggests that the same possibilities exist for automated bibliographic control. Archivists should not let similar possibilities for economy of scale pass them by. Doralyn J. Hickey, writing in *Literary Trends* (July 1976) indicated that initiative in devising subject bibliographic control had passed largely to the information specialists. This need not remain so nor need it apply to archives.

Dick Lytle's work can be a significant first step in the development of more precise, more comprehensive, and more cost effective systems which use automation. His study needs to be expanded for many reasons. I know, speaking from the point of view of BRISC alone, that additional questions are urgent:

—how satisfied are various kinds of users? (institutional, scholarly, administrative, etc.)
—how competent a job do staff feel they perform?
—what is the average "unit cost" for each search?
—how do staff perceptions of performances compare with user perception?

W. Theodore Dürr, *Director*
Baltimore Region Institutional Studies Center