Published in: *Information Processing and Management* **28, 457-466 (1992).**
Published version

On the evaluation of IR systems

S.E. Robertson and M M Hancock-Beaulieu

Centre for Interactive Systems Research
Department of Information Science
City University
Northampton Square
London EC1V OHB, U.K.

Final version 4 December 1991

Abstract

The paper highlights the ever increasing complexity in the evaluation of IR systems which has arisen over the last decade. Relevance, cognition, user behaviour, interaction and a changing view of the boundaries of the system are considered to be contributory factors. Issues such as laboratory versus operational systems, black-box versus diagnostic experiments, and qualitative and quantitative methods, are discussed and supported by examples drawn from three groups of evaluative experiments: weighted searching on a front end system, information seeking behaviour and the use of OPACs, and the OKAPI experimental retrieval system.

The volume edited by Sparck Jones and published in 1981, *Information Retrieval Experiment*, remains the one substantial work on the evaluation of IR systems. The first chapter, by one of the present authors (Robertson, 1981), ended with the thought that the succeeding twenty years might see as much change in this field as the previous twenty:

> If, in 2001, this entire chapter is obsolete, so much the better!

As we are now half way through that period, an overview is appropriate. It is the contention of the present paper that the field has changed substantially in ten years.

1 INTRODUCTION

Some of the issues to be explored in this paper may be briefly previewed.

One issue discussed in Sparck Jones (1981) is the idea of laboratory versus operational system tests. This remains a difficult issue, indeed in some ways has become more difficult, in that some of the research questions one would now like to be able to explore in a laboratory environment are less suitable for that environment than was the case ten or twenty years ago. This reflects some deeper question which are explored here.

Secondly, an implicit issue in the book is that of black-box versus diagnostic experiments. (A black box experiment is one in which the system is treated as a whole, and inputs are controlled, outputs observed and evaluated. A diagnostic experiment is one in which it is not so much the outputs themselves that are of interest, as why they occur, in terms of the

internal structure, components and activities of the system.) Once again, it is argued that black-box experiments are becoming inherently more difficult to perform, given the kinds of research question currently at issue, and that this problem is a reflection of deeper questions.

Thirdly, there is a possible distinction between qualitative and quantitative methods. Actually, this is something of a false dichotomy, in that all IR experiments involve qualitative methods and most also involve quantitative methods. The problems arise in deciding what kind and types of qualitative judgments to make or include in an experiment, and here again the choice may have become more difficult.

The final issue in this list, but the one with which we start below, is that of identifying the boundaries of the system(s) to be experimented upon. Herein, indeed, lie some of the deeper questions referred to above.

As we explore these issues and questions, we will use as illustration a number of experiments recently conducted at City and elsewhere. The three groups of experiments are:

1. An evaluation of weighting, ranking and relevance feedback via the front-end system Cirt;
2. A series of experiments on the OKAPI experimental retrieval system (begun at the Polytechnic of Central London and continued at City); and
3. A series of experiments concerning OPACs and the information-seeking behaviour of library users.

2 SYSTEMS

A systemic approach to information retrieval might start from the question of where the boundaries of the IR system lie. In general terms, any (open) system has interactions with its environment, which might, depending on the context, be termed inputs and outputs or perceptions and responses. (The environment of a system is of course the remainder of the universe). Thus we have the following:

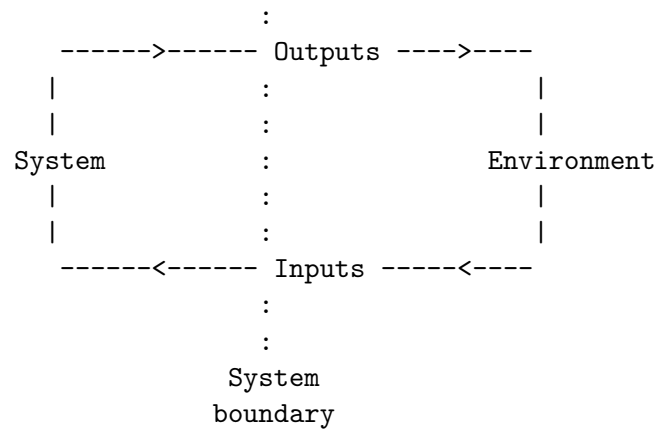

Fig. 1. An open system.

In the early days of IR system tests, the interpretation of this diagram was fairly straightforward. The "system" was the retrieval mechanism and associated human activities (indexing, searching, etc); input was the stated request, and output the retrieved items. Evaluation

consisted in assessing each retrieved item against the input request, the assessment being made by an independent judge. By the same token, it was easy to regard the system as a black box, of which only the inputs and outputs were visible.

This was exactly the approach taken, for example, in the first Cranfield experiment (Cleverdon, 1962). Each of the four systems (including, of course, the experts to run it) was regarded as indivisible, and the emphasis was on testing the entire system. Independent judges assessed the output against the input.

However, much of the development of ideas in IR research in the last thirty years has been in directions which undermine this simple model. Although we may still be looking at systems in the broad sense, the boundaries of those systems become progressively harder to define. There are perhaps three major components to this process.

2.1 The relevance revolution

Relevance is the most highly debated aspect of IR research, and it would be foolhardy to claim that there exists any kind of paradigm or consensus on the subject. Nevertheless, starting from Taylor's classic paper (Taylor, 1968), there has been increasing acceptance that stated requests are not the same as information needs, and that consequently relevance should be judged in relation to needs rather than stated requests. (A variant on this theme requires that relevance should be observed behaviourally, i.e. should be inferred from some action on the part of the requester.)

Where does this leave the boundary of the "system"? Apparently we must include within the boundaries either some self-reported mental activities, or some externally observable activities of the user. In other words, the "system" for evaluation purposes is no longer identified with the IR system in the traditional sense. For this reason, we prefer the term "mechanism" for the latter, narrow system.[1]

2.2 The cognitive revolution

By an extension of the same process, we have come to see information seeking in cognitive terms. This is best illustrated by the ASK idea (Belkin, 1980), in which the information need is seen as a reflection of an anomalous state of knowledge on the part of the requester.

From this point of view, the system (for evaluation purposes) might be seen as including all the user's activities while interacting with the mechanism. Thus the input is a user with a certain (anomalous) state of knowledge, and the output is the same user with an altered state of knowledge (anomaly resolved or whatever).

2.3 The interactive revolution

At the same time as these conceptual developments were taking place, IR systems (mechanisms) were becoming more (or more obviously) interactive. Thus the image of feeding in a question and getting out an answer seems out of place. Instead, it appears that in order to make any sense at all of the differences between rival mechanisms, we need to look carefully at the process of interaction.

[1]The term "mechanism" was used in this sense by B.C.Brookes, who died during the gestation of the present paper, and to whose memory it is hereby dedicated.

The problem becomes very much more serious when we consider the time factor. Much of the early testing was conducted in an environment where putting a question to a system was a major event (e.g. writing to the National Library of Medicine and waiting for a reply - Lancaster, 1968). Given the increasing availability of easily accessible systems, it is entirely feasible for an individual to use a mechanism repeatedly, on essentially the same topic, as his/her state of knowledge develops or changes. (They may also, of course, interleave uses of the mechanism in question with other formal or informal mechanisms, including for example discussion with colleagues.) Thus the resolution of an ASK can no longer (if indeed it ever could) be equated with a single search session.

These three sea-changes in our perception of retrieval have profoundly influenced the nature of the evaluation problem. Some effects on the issues identified in the introduction are discussed below in the context of particular experiments.

3 SOME EXPERIMENTS

The three groups of experiments discussed in this section are all described more fully elsewhere. The purpose of the present discussion is to emphasise and exemplify the issues introduced above. Each group of experiments will be identified with a particular issue, not because the other issues do not arise, but because the experiment illustrates that particular issue well.

3.1 Black-box versus diagnostic experiments

The classic diagnostic experiment was the Medlars evaluation of mid-sixties (Lancaster, 1968). Although this experiment involved both relevance judgements and recall/precision calculations, the latter were not the most important part of the analysis. The relevance judgements were used to identify and categorise failures, in terms of system features such as the index language or search strategies. Thus the experiment led directly to recommendations for improving performance by modifying the internal mechanisms. Most experiments involving test collections, on the other hand, take a black-box form. (By conducting a large number of black-box experiments, on permutations of variables, one might be able to make diagnostic inferences, and this is often the intention of test-collection experiments.)

One might look to a black-box experiment, either to decide between two or more competing systems (e.g. two databases on the same host, where the experimenter does not have control over the selection or indexing policies), or to decide between competing principles or approaches (e.g. partial match versus Boolean).

In order to undertake a black-box experiment, one must have clearly defined system boundaries, inputs and outputs. Furthermore, the inputs and outputs must in some sense be observable or measurable. Clearly, states of knowledge (whether anomalous or otherwise) are not directly observable; much of the difficulty of conducting black-box experiments with our present perspective lies in this fact. Experimental designs must be concerned with methods of observing inputs and outputs, and with operational definitions of the variables concerned.

3.1.1 The Cirt evaluation project

The Cirt project is an example of an operational evaluation which comes nearest to being a black-box experiment. The object of this experiment (Robertson and Thompson, 1990) was

to evaluate, in an operational environment, Weighted against Boolean searching. ("Weighted" searching included search term weighting, ranking of retrieved references and relevance feedback, and "Boolean" was shorthand for the methods traditionally used on commercial hosts, including Boolean and pseudo-Boolean operators and intermediate search sets). It was in essence a black-box experiment, in that it was not concerned primarily or directly with why one system performed better or worse than another (although some such hypotheses were considered later). It followed a series of laboratory experiments on Weighted searching, by a number of researchers.

The experimental design actually dated from 1979, in the form of an unpublished proposal to the British Library by Jamieson and Oddy. For various technical reasons, their project was never completed, but the ideas were revived for the Cirt experiment, which took place between 1985 and 1987.

The idea was to catch users approaching an information retrieval service, and assign them at random to either Boolean or Weighted searching (an independent sample design). The system in this case consisted mainly of a two-part mechanism (traditional Boolean host plus front-end to allow Weighted searching), and also a trained and experienced intermediary. The user was part of the system at least inasmuch as s/he was interviewed by the intermediary, was present at the search (primarily to make online relevance judgements as part of the feedback process), and received the printout later for the evaluation relevance judgements. The intermediary could be expected to interact with the user in whatever way s/he saw fit, so in practice the user's contribution may have been somewhat greater. Indeed, it was the perception of the cognitive nature of this interaction which forced the independent-sample design mentioned above. However, the matter was treated very much as a clearly delineated process, starting when the user approached the information service and ending with the supply of printout lists. In this sense, it fits well with the 1981 model.

The evaluation was based on traditional relevance criteria, and also on various (mainly qualitative) assessments at the question level. These latter assessments were provided both by the users and by the intermediaries, and concerned general satisfaction with the search, amount of effort involved, perceived difficulty, time taken etc.

The overall conclusion was that Weighted searching was comparable to Boolean. However (this is where it becomes more diagnostic), some specific limitations of Weighted searching in that kind of environment were identified. The major experimental problem was that of obtaining sufficient requests/users for experimental purposes under the independent-sample design (Robertson, 1990). Thus the beginnings of the cognitive revolution referred to in section 2.2 led directly to problems in the quantitative aspects of the experiment. This problem, in turn, led us to look for an experimental situation that would allow larger numbers of users to be studied. The OPAC environment, the subject of section 3.3.1, was one such.

3.2 Laboratory versus operational experiments

The conflict between laboratory and operational experiments is essentially a conflict between, on the one hand, control over experimental variables, observability, and repeatability, and on the other hand, realism. In the early days of testing, the realism question centred on the requests and relevance judgements: as testers became aware of the relevance revolution, it became clear that a good treatment of the relevance issue required real people with real information needs. This factor remains a powerful argument in favour of operational tests, but other factors have added arguments on the same side. For example, it is extremely

difficult, maybe impossible, to design a reasonable laboratory test of a highly interactive system, simply because we do not know how to simulate a real user's reactions. On the other hand, operational testing is not easy either.

The ideal combination of laboratory and operational tests would presumably start in the laboratory, investigating combinations of factors under controlled conditions, and move towards operational tests on a much smaller set of options. The OKAPI work illustrates this principle. Also the Cirt evaluation represents the operational phase of a series of experiments on probabilistic methods. But even within a given experiment, it will be clear that there is frequently conflict between the two principles, and any experimental design must be a compromise.

3.2.1 Early OKAPI experiments

OKAPI is a third generation, experimental online catalogue first created in 1985. Since then the system has been used as a testbed to explore and evaluate different approaches and mechanisms to improve retrieval. Several successive versions of the system have been developed. The design philosophy has been to build a system for naive users and to test the different prototypes in a live environment. As the design cycle has progressed, the evaluative experiments have also developed to combine both laboratory and operational approaches.

The first experiment was to test a best match system which included weighted searches and ranked output (Mitev *et al*, 1985). The initial evaluation was primarily based on an analysis of the transaction logs of searches carried out by users of the system installed in a library. In addition reactions from users on the acceptability of the system were also sought through some brief post-search interviews. The log analysis produced diagnostic data which provided some clues for improving subject retrieval. This then formed the basis for the next step in the system development.

In the second experiment new retrieval mechanisms were introduced to improve recall (Walker and Jones, 1987). These included stemming, automatic spelling correction and cross-reference tables. To evaluate these features a controlled experiment was set up using different versions of the system. Two versions were installed in the library in different locations. One contained all the devices, including weak and strong stemming, and the other included weak stemming only, i.e. plurals, '-ing' and '-ed' endings. Since the searching features were transparent, users were not aware of any difference between the systems available.

A third version contained none of the new retrieval aids and served as a control. This was used to repeat searches which had been identified from the transaction logs of the library versions. The experimenter selected those searches where there was some degree of confidence that they could be repeated realistically. Searches conducted on one of the library versions could also be repeated on the other. Thus even in an operational setting an attempt was made to control the variables so that the retrieval effectiveness of the different devices could be compared.

In addition to the retrieval test, a second study was undertaken to assess users' opinions of the improved system. Users were asked to compare between the experimental system and the library's commercial online catalogue (Jones, 1988). This served to register users' satisfaction and preferences for the different features of the two systems. Moreover it confirmed that users' perception of the performance of the experimental system matched the system's actual performance as established in the formal experiment.

3.2.2 Recent work on OKAPI

The third experiment was concerned with the evaluation of a highly interactive system which included relevance feedback and query expansion (Walker and DeVere, 1990). The evaluation was carried out in two stages, a laboratory test followed by an operational one. In the controlled laboratory experiment some attempt was made to select subjects who were representative of the actual user population. Subjects were given the task of drawing up a reading list for a set of topics. Each user chose one topic and carried out a different search on each of the two versions of the system. One version included the automatic query expansion feature, the other had both query expansion and a facility to browse items with the same classification number as a retrieved item. A brief post-search interview also sought users' opinions.

In a somewhat similar fashion to the Cirt experiment described above, relevance judgements were made at two stages in the experiment: by the searcher in the course of interacting with the system, and subsequently by a panel of judges. However, unlike the Cirt evaluation, the diagnostic nature of this experiment required that the online judgements should also be used as part of the evaluation. A device or feature would be useful if it led the searcher to items that s/he deemed relevant, even if the panel disagreed. Furthermore, because these searcher judgements would not necessarily be consistent between different searchers, it was appropriate for the experimenter to make qualitative assessments at the topic level on the basis of the searcher judgements (e.g. whether performance of the device on this particular search was good, moderate or bad), and then to summarise these qualitative assessments over topics. The searcher judgements also had to be compared with panel judgements.

On the basis of the results of the laboratory tests, only the query expansion facility (in addition to the best match and the other searching aids previously tested) was included for the operational test (Walker and Hancock-Beaulieu, 1991). The system was installed over eight months in the library and was also accessible over a network together with two other bibliographic databases which were also mounted with OKAPI search software.

One problem which has bedevilled OPAC research based on transaction logs is the demarcation of search sessions. An individual sitting at a terminal may restart his/her search several times on the same problem or closely related ones, or may conduct several quite distinct searches. This is a very practical problem for the experimenter, in that many logging systems provide no mechanism for identifying users (in fact OKAPI requires a log-on process). But it is also a problem of principle for the design of evaluation experiments, in that it introduces another level into the evaluation: instead of document/query we have document/search-statement/session, with an implicit need level which may be different again.

For example, in this case a search-level analysis suggested that on most occasions, query expansion was not used. In interviews (session-level), many users indicated that they did not intend to undertake exhaustive searches and that the best match search produced adequate initial results. A few users were also not aware of the availability of the option.

A select number of searches in which the query expansion was not used, was repeated by the experimenter to include query expansion. In 50% of the cases, new references which were judged to be relevant were retrieved. The replay of searches in the same way, in the presence of searchers, also produced a similar result. The long term monitoring of the usage of the system also provided the opportunity to observe changes in searching behaviour. For example, first time users as well as more experienced users of the system were found to use query expansion more frequently, whereas intermediate users expanded their searches less often.

Further work is under way concerning the continuity of search topics and relevance judgements over successive sessions by the same user. Preliminary results suggest a surprising number of instances of identical or closely related searches by the same user at intervals. This suggests a user exploring a topic over an extended period of time, and seems to demand yet another level of analysis (beyond search-level and session-level), namely problem-level. To our knowledge very little attempt has been made to evaluate systems at the problem-level, as represented by information-seeking behaviour over an extended period (exceptions are Ellis, 1989, Kuhlthau, 1991 and Smithson, 1990).

3.3 Qualitative and quantitative methods

The traditional, Cranfield-like, retrieval experiment appears at first sight to be quantitative in emphasis. However, such a view ignores the very important qualitative aspects. In effect, a traditional comparison involves the following stages:

1. Qualitative assessment at the level of question-document pairs, of relevance
2. Quantitative analysis covering both the different documents and the different questions
3. A final qualitative assessment of which system(s) perform better than other(s)

In other words, the qualitative aspects are confined to the two ends of the process of assessment.

However, even within that tradition, there are evaluation experiments which introduce other qualitative aspects. One example comes out of the apparently very quantitative problem of significance testing. One test which has been used in some experiments (Salton and Lesk, 1968) is the sign test. This involves an intermediate qualitative stage in the above process, where data concerning a single question but all documents is reduced to a qualitative form (system A better than/worse than/the same as system B) before cumulating over questions. Although we are not aware of any instance, the same test could be applied if the original qualitative assessments were at the question level instead of at the lower level of question-document pairs. For example, users might be asked to make comparative judgements on systems after performing the same search twice, and the results subjected to the sign test. More generally, almost any experiment in IR contains some qualitative and some quantitative elements. For example, the Medlars experiment involved the (qualitative) classification of failures into categories, whose statistical prevalence was then measured.

The OPAC evaluation project, described below, concentrated on qualitative assessments (by the experimenter and/or the subjects) of the users' information-seeking activities and their perceptions of those activities. These qualitative judgements were then subject to quantitative analysis by cumulating over users. The Cirt evaluation involved relatively traditional relevance assessments as well as some request-based qualitative judgements.

A critical design decision for today's experimenter concerns the nature of the qualitative assessments that are needed in order to address the particular research questions under investigation. Certainly the quantitative emphasis in some experiments should not be taken as an indication of objectivity, validity, or reliability simply on the basis of its quantitative nature. Furthermore, the easy option for the laboratory researcher, of using someone else's relevance judgements as embodied in a test collection, carries no guarantee of answering the right questions.

3.3.1 Opac evaluation project

The OPAC Evaluation Project at City University (Hancock-Beaulieu, 1990; Hancock-Beaulieu *et al*, 1991) was concerned with the use of the library catalogue in the broader context of user information seeking behaviour. In one experiment searching activity at the catalogue and/or at the shelves was observed for both catalogue users and non-users. The system boundaries thus encompassed several aspects of user behaviour and the library environment as a whole. In essence the user was regarded as part of the system for an interval of time; thus the input might be defined as user-with-information-need, and output as user-with-books. In this respect it was not very different in approach from the Cirt experiment, where we could have defined input and output similarly. But in the OPAC studies the emphasis was very much on the intermediate processes, rather than on input and output.

It was not feasible to adopt a holistic approach in all circumstances. In other experiments users of the catalogue were not followed to the shelves but were asked what they intended in the use of the tool as well as to indicate how they would proceed as a result of the catalogue consultation. The search outcome was then assessed according to intention and results. The discrepancy between what was observable through the transaction logs and users' declared intentions confirmed the importance of correlating the two sources of data to obtain more reliable evidence on information seeking behaviour and search outcome.

A combination of data gathering methods was used to obtain both quantitative and qualitative data about the entire search process, eg. observation, talk aloud, questionnaires and transaction logs. The quantitative data mapped out overall search patterns and strategies. The main focus however was on obtaining qualitative data on how searches evolved by comparing within each individual search session: how users first articulated their information need; then how they formulated and reformulated their search in the catalogue; and finally where applicable, how the documents selected on the shelves reflected the articulated need and search formulations/reformulations.

In analyzing the different transitions in the course of a search, some attempt was made to understand the interactive nature of the retrieval task. The main problem is to determine to what extent any particular change or development in the search formulation stems from a positive influence of the system, or is due to the searcher adapting to the system. Alternatively it could also be a result of a genuine change in the searcher's state of knowledge or the searcher expressing an existing but previously unexpressed need. In reality it may well be a combination of these factors.

The experiments undertaken seem to indicate that a combination of data gathering methods and techniques are necessary to elicit information from users. Attempts to encourage users to be more discursive about their information needs were not very successful. There are obviously limitations in trying to extract explanatory information from the user in the course of a real search. Although users did respond to online in-search questions, seemingly without having the interruption interfere with the on-going search, these questions were not probing.

A further method of data-gathering, which was tried in embryo only but showed some promise, was to allow the user to complete the search uninterrupted, but immediately to replay the search on the screen and invite the user to provide some diagnostic feedback. The replay facility simulated the real time of the search, so that the user could see immediately where they had paused or moved on rapidly.

4 DISCUSSION AND CONCLUSIONS

4.1 Mechanisms and behaviour

The discussion in section 2 on the identification of system boundaries suggested that we are moving in the direction of a broad view of IR "systems" encompassing human activities well away from the mechanism. The experiments described in section 3 have illustrated the nature of this process and confirmed the trend, and two of the main issues identified (black-box versus diagnostic and laboratory versus operational) may be seen as direct reflections of this central question. Given that in the past most experiments have been essentially black-box and laboratory-based, our conclusion from this analysis is that more emphasis should be placed on diagnostic and/or operational experiments.

The third issue, qualitative and quantitative methods, is also associated with this central question, in that the system boundary is the focus of most measurement/assessment processes, and the nature of these processes is strongly constrained by the location of the boundary. Diagnostic experiments also require measurements or assessments away from the boundary, but very often the assessments required are essentially qualitative.

Despite this general widening of our evaluation horizons, in many experiments the focus will remain the mechanism (being that part of the system over which the designer or manager has control). However, in some experiments, the emphasis will no longer be on the mechanisms at all, but on human behaviour. Indeed, we may see that studies right outside the framework that we associate with IR system evaluation become, in retrospect, relevant to our present concerns. Thus for example the work of TJ Allen (Allen, 1968), on individuals' perceptions of the sources of information available to them, and on their use of these sources in the context of their needs, addresses what must now be seen as a crucial aspect of IR system evaluation.

4.2 Methods for interactive systems

While the interactive revolution has contributed greatly to the difficulty and complexity of undertaking IR system evaluation experiments, it also has a more positive side. We have unprecedented opportunities for developing tools and techniques in the conduct of experiments which have the potential to increase vastly our knowledge of human information-seeking activities.

We may start with the simple transaction log. While the usefulness of this device has long been recognised (eg Rice and Borgman, 1983), it has also been limited by a number of factors:

1. Logs usually hold information about the use of commands only, not about system response;

2. It is usually difficult to demarcate sessions at public terminals, in cases where the system has no log-on procedure;

3. Straight transaction logs provide information only about what the user did, not what s/he thought.

These are indeed considerable limitations. However, none of these problems is an absolute barrier.

1. *Data held in log*: Although few commercial systems provide such a facility, there is no difficulty in principle in logging everything that takes place in a session. Clearly such logs are voluminous; it is then necessary to summarise the logs in various ways. However, some investigations can take place with the researcher replaying logs in full, and noting events which require a human interpretation.

2. *Session demarcation*: Although some OPAC providers have objections in principle to making the user log on, there are many other environments where that is not a problem. There will of course be cases where one user finds the terminal still logged on after the previous use, but an automatic time-out will reduce this problem.

3. *Enhancing logs*: The feasibility of enhancing transaction logs with additional information requested from the user has now been demonstrated. Enhancement can take the form of questionnaires of various kinds administered automatically by the system before, during or after a search, or of the researcher interviewing the user at some stage, perhaps immediately afterwards with a replay of the log to the user.

4.3 Test collections versus evaluation facilities

Much traditional work on IR system evaluation has taken place using one or more of the traditional test collections. The existence of portable test collections (with queries and relevance judgements) has been a substantial factor in the development of research in the field. Operational tests have often been one-off exercises, strongly directed at the needs of a particular organisation or service.

The pressures that have been discussed in this paper, in the direction of more diagnostic and more operational system tests, seem to encourage the development of an intermediate stage: the evaluation facility. The characteristics of such a facility would be: an operational environment, with real users with real, live problems; a live database (or bases) with the capability to serve at least some of the needs of these users; a fully functional but experimental system (mechanism), capable of being easily modified; and a range of data-gathering and analysis tools for evaluation purposes.

Such facilities would complement the existing, well-established, test-collection-based evaluation paradigm, and greatly extend the range of possible experiments. A few facilities with some of these characteristics do exist; the operational experiment discussed in section 3.2.2 was made possible by the development at City of an evaluation facility based on OKAPI. We will shortly be embarking on some experiments in which users of OKAPI may be invited during the search to interact with some thesaurus-based knowledge structure – e.g. to replace or add to an initial query with terms taken by the system from a thesaurus. This will no longer be a transparent mechanism (as are most of OKAPI's present features), and thus will take us further into research areas for which test collections would be inadequate.

Clearly the development of more effective IR systems (mechanisms) is dependent on the development of effective methods of evaluation. The challenge for the next decade is to explore the multiple dimensions and components of the new generation of information retrieval systems by experimenting with a diversity of evaluative approaches.

Acknowledgement – We would like to thank Donna Harman and a referee for valuable comments on an earlier draft of the paper.

REFERENCES

Allen, T.J. (1968). Organisational aspects of information flow in technology. *Aslib Proceedings*, 20, 433-454.

Belkin, N.J. (1980). Anomalous states of knowledge as the basis for information retrieval. *Canadian Journal of Information Science*, 5, 133-143.

Cleverdon, C.W. (1962). *Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems*, Cranfield: College of Aeronautics.

Ellis, D. (1989). A behavioural approach to information retrieval system design. *Journal of Documentation*, 45, 171-212.

Hancock-Beaulieu, M. (1990). Evaluating the impact of an online library catalogue on subject searching behaviour at the catalogue and at the shelves. *Journal of Documentation*, 46, 318-338.

Hancock-Beaulieu, M., Robertson, S.E. & Neilson, C. (1991). Evaluation of online catalogues: eliciting information from the user. *Information Processing and Management*, 27, 523-532.

Jones, R.M. (1988). *A comparative evaluation of two online public access catalogues.* London: British Library.

Kuhlthau, C.C. (1991). Inside the search process: information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42, 361-371.

Lancaster, F.W. (1968). *Evaluation of the MEDLARS demand search service*, Bethesda, Maryland: National Library of Medicine.

Mitev, N.N., Venner, G.M. & Walker, S. (1985). *Designing an online public access catalogue: OKAPI, a catalogue on a local area network.* London: British Library.

Rice, R.E. & Borgman, C.L. (1983). The use of computer monitored data in information science and communication research. *Journal of the American Sociey for Information Science*, 34, 247-256.

Robertson S.E. (1981). The methodology of information retrieval experiment. In: Sparck Jones, K. (Ed.), *Information retrieval experiment.* London: Butterworths, 9-31

Robertson, S.E. (1990). On sample sizes for non-matched-pair IR experiments. *Information Processing and Management*, 26, 739-753.

Robertson, S.E. & Thompson, C. L. (1990). Weighted searching: the CIRT experiment. In: Jones, K. P. (Ed.), *Informatics 10 – Prospects for intelligent retrieval.* London: Aslib, 153-165.

Salton, G. & Lesk, M.E. (1968). Comparative evaluation of indexing and text processing. *Journal of the Association for Computing Machinery*, 15, 8-36.

Smithson, S. (1990). The evaluation of information retrieval systems: a case study approach. In: Jones, K.P. (Ed) *Informatics 10 – Prospects for intelligent retrieval.* London: Aslib, 75-89.

Sparck Jones, K. (1981). *Information retrieval experiment.* London: Butterworths, 1981

Taylor, R.S. (1968). Question negotiation and information seeking in libraries. *College and Research Libraries*, 29, 178-194.

Walker, S. & Jones, R.M. (1987). *Improving subject retrieval in online catalogues. 1. Stemming, automatic spelling correction and cross-reference tables.* London: British Library.

Walker, S. & DeVere, R. (1990). *Improving subject retrieval in online catalogues. 2. Relevance feedback and query expansion.* London: British Library.

Walker, S. & Hancock-Beaulieu, M. (1991). *OKAPI at City: An evaluation facility for interactive IR.* London: British Library. (British Library Research Report no. 6056).