Chapter 2

Approaches to Evaluation for Library Catalogues

Introduction

This chapter presents a review of the different approaches to evaluation and in particular the data gathering methods which have been used in both traditional library catalogue use studies and in online catalogue research. It is suggested that the methodological shortcomings of the quantitative surveys in the evaluation of manual catalogues may have led to an incomplete picture of the search process and possible misinterpretation of user needs. Online catalogue research has adopted a diversity of methods including comparative studies, prototyping, controlled experiments, transaction log analysis and protocol analysis. These take into account system performance through retrieval tests and user performance through diagnostic analysis.

2.1 Evaluation and traditional library catalogues

The main objective of traditional catalogue use studies has been to collect data on the use or usage of the tool and then by inference to draw conclusions about user behaviour and user needs. The emphasis appears to have thus been placed on the means rather than on the end, i.e. the tool as opposed to the task. A closer examination of the data gathering methods applied illustrates some of the methodological shortcomings of these studies.

2.1.1 Questionnaires and interviews

In traditional catalogue use studies, the survey method including questionnaires and interviews has been the dominant form of data gathering. The point at which the searcher was interrogated has varied and has to a large extent determined the type, amount and reliability of the data collected.

In the U.K. catalogue use survey (Maltby, 1973), library users were interviewed as they left the library and were asked to recall their last search as well as to generalise on how they would normally search. Lipetz (1972) in the Yale study, interviewed users

before and after they had consulted the catalogue and focused on a more immediate and specific search. In doing so he was able to ascertain to some extent, whether or not users really did what they set out to do and thus discovered that searchers' immediate and underlying objectives differed.

Tagliacozzo et al (1970) went further by observing and questioning subjects about success and failure at different stages during their consultation. By identifying the different strategies adopted by searchers after their initial access to the catalogue, the progression of the search was thus partly followed. This was achieved by the experimenter noting if the user moved from one entry point of the card catalogue to another, whereupon questions could then be asked about the different access points. The catalogue consultation was also followed up by questioning users when they returned from the shelves and it was found that subsequent shelf-browsing was widespread.

By their very large scale, these studies aimed to gather quantitative data on use and by the very nature and breadth of the samples (over 2,000 cases), in-depth qualitative analysis of user activity was limited. In not observing the search at first hand, the indirect and partial approach resulted in an incomplete picture of catalogue searches. Consequently, little is revealed about searching as a process and the interactive nature of user searching behaviour at the catalogue.

2.1.2 Protocol analysis

Experimenters avoided a more direct approach to eliciting information from the user on the grounds that it would interfere and possibly distort the search process. Ericsson and Simon's (1980) seminal paper on verbal reports as data has led to the use of protocols or spoken thoughts in task analysis and decision making processes in a number of other areas.

Markey (1984) applied protocol analysis to study manual subject searching at the library catalogue. With minimal prompting by the experimenter, the searcher was encouraged to talk aloud as the search was carried out. The proceedings were recorded on a tape recorder and from the transcript a flow chart of the different steps in the complete search was produced, indicating all the decision points and actions taken by the searcher. From the coded protocols, searching patterns of behaviour emerged and models of different types of searches were identified. It was found that the majority of searchers extracted a class of numbers from the catalogue and proceeded to continue their search at the shelves rather than at the catalogue. This method of eliciting information from the user does generate a vast amount of data. It appears to be an effective way of obtaining a typography of searches but the transcription and subsequent coding are very cumbersome and time consuming.

To circumvent some of the disadvantages of protocol analysis Hancock (1987) devised a combined observation and talk aloud technique to study searchers using a microfiche catalogue and printed subject index. Data was recorded on a highly structured dual purpose observation and questionnaire form incorporating some real time interpretation. Subjects were encouraged to talk aloud as they searched merely to give an indication of what they were looking for and to confirm whether or not they

succeeded.

The direct observation of searchers' actions provided the experimenter with the framework of the activity whilst the verbal data gave the details which could not be easily observed or which were not observable. Searchers' actions thus confirmed what they said they were doing or vice versa. In order to get a more complete picture of the information seeking behaviour, a holistic approach was also adopted by including searching activity at the shelves following the catalogue consultation.

2.1.3 Summary of methodological shortcomings

Although the limitations in the performance of traditional library catalogues particularly for subject searching, had been recognized, (Atherton, 1978), Hafter (1979) in her review of catalogue use studies states:

> The major conclusion that emerges from these studies is that the card catalog works. Even more importantly, users are skillful at manipulating it for their own purpose.

The general results of low usage and the dominance of the catalogue as a finding tool for specific item searching does not appear to support this conclusion. It could be argued that the apparent preference for specific item searching may have been more a function of the catalogue's limited subject access capability than a reflection of the user's information need. The very design of the tool could well have been encouraging a certain type of behaviour, that is the tool tailoring the task.

In the study of users and the usage of the library catalogue, a number of difficulties and constraints have been met in data gathering, namely:

- Observation of the user is difficult, the physical format being only one of the obstacles.
- The catalogue consultation is not a static process but consists of a series of events and the searcher's pattern of behaviour depends on what comes before and after the different stages. The links between each step of the searching process are just as much an integral part of the search as the individual steps.
- The structure of the catalogue influences user behaviour and does not necessarily reflect information needs, i.e. user performance is dependent on systems performance.

Moreover some of the methodological limitations have also stemmed from:

- a dependence on the catalogue consultation alone to inform on the information seeking process as a whole,
- a reluctance to use more direct or diverse means of eliciting information from users for fear of interfering with the search process,
- a reliance on users themselves to provide adequate information before and after the event.

2.2 Experimental and analytical methods for OPAC evaluation

Whereas traditional catalogue use studies have been dependent on the survey method and based on collecting data from the user in an operational setting, online catalogue studies have adopted a more experimental and controlled approach. Partly because of the technological environment, a number of different methods and combinations of methods for data gathering and evaluation have been used.

These can be divided into two groups. The first consists of more general testing methods which aim to measure system performance through retrieval tests, and include comparative studies, prototyping and laboratory type controlled experiments. The second group are more user orientated, diagnostic analytical methods, such as transaction log analysis, protocol analysis and talk-aloud techniques, as well as the other usual methods of eliciting information from users i.e. questionnaires and interviews. The design of the evaluative component of an experimental study could include a combination of methods from both groups. We shall draw on the major online catalogue studies for a critical assessment of how each of these two categories of evaluative methods have been applied.

2.2.1 Comparative studies

Several different types of comparative studies have been undertaken. Gouke and Pease (1982) compared searching for assigned titles on an online catalogue as well as on a card catalogue. The object was to assess user acceptability of the new medium. In spite of limitations in the performance of the online catalogue, users nevertheless expressed their preference for the computerized tool.

The trial test of Cite, the online catalogue at the National Library of Medicine, had a much more system defined objective in comparing the performance of two online catalogues, (Siegel et al, 1984). In an attempt to differentiate system dependent and independent variables, a combined methodology was devised. Two separate groups of library staff and users were assigned searches on either system. In addition to this 'sample search experiment', a 'comparison search experiment' was also conducted whereby users searched both systems for topics of their choice. A survey using a self-administered questionnaire provided additional user data.

The main drawback in this type of general comparison in an operational setting is the inability to control the variables across the two systems. We are not comparing like with like. In this case the searching capability of Cite was far superior to that of its rival and so it was not surprising that it performed better and users preferred it. It may have been more valuable if it had been possible, to have tested the contribution of individual features to the overall result.

A similar disadvantage is to be found when Okapi'86 was compared with the operational system Libertas to determine a) user preference for specific features, b) user assessment of performance and c) user attitudes to Okapi's recall improvement devices, (Jones, 1988). In this case however, evidence could be corroborated with

other Okapi evaluative studies, (Walker & Jones, 1987). The quantitative measure of recall was also correlated with the qualitative measure of user satisfaction. Although satisfaction was measured in terms of number of references, there may have been other contributory factors which were not accounted for.

In all of these comparative studies although system performance is the dominant interest, there is nevertheless some attempt to take the user into account to some extent particularly in the Okapi'86 evaluation.

2.2.2 Prototyping

Prototyping as a methodology can be an effective way of developing and testing individual system design features. This is not an uncommon method in numerous other computer applications for example in manufacturing and processing. Unfortunately this method as an evaluative method has only been used in the design of experimental systems. In spite of the many in-house online catalogues and the commercial systems available, evaluative data on their different stages of development, if it does exist, has not been made available.

The Okapi'84-86 projects funded by British Library are a major exponent of this approach, (Mitev, Venner and Walker, 1985, Walker and Jones, 1987). Each successive version built on the results of the former. Evaluation was carried out in an operational setting with real users and real searches. Failure analysis of transactions logs of the first version Okapi'84 led to improvements of the combinational search mechanisms for partial matches and new devices to improve recall in the second version Okapi'86. These included automatic stemming or truncation, cross-reference tables and spelling corrections.

To evaluate these features three catalogues were used. EXP contained all the devices, including both weak stemming (i.e. plurals, ing and ed endings) and strong stemming (i.e. tion, ness, ist endings) and CTL included weak stemming only. The third catalogue, OSTEM, contained none of the new retrieval aids and was used as a control to repeat searches which had been identified from the transaction logs of the other two catalogues in the library trials. Thus by isolating variables the retrieval effectiveness of the different devices was tested and compared. 50% of spelling errors were corrected with favourable results and weak stemming was found to increase recall without affecting precision. Strong stemming on the other hand was more tenuous.

The Dewey Decimal Classification Online Project (Markey & Demeyer, 1986) also produced a prototype system which was evaluated in four libraries in the form of two catalogues featuring different subject searching capabilities. One included keyword access in titles and LCSH headings and the other was enhanced with the DDC classification schedules and the relative index. Retrieval tests were conducted using comparative as well as sample searches as with the CITE experiment and were followed by a post-search interview which provided some insight into user expectations and behaviour.

Comparative results of estimated recall and precision based on references displayed, showed that the more traditional keyword access performed better than the enhanced system, however the latter retrieved a different set of references. Moreover failure

analysis revealed that the Dewey catalogue provided more relevant items for searches which were dependent on user-entered terms. At the same time success was not dependent on user-entered terms alone in that the user also had the opportunity to further specify and search. Improvement in subject access can thus be seen in terms of quality and not simply quantity. Recall and precision alone do not appear to be adequate measures of success.

In addition searchers expressed difficulties in choosing between all the different options. Whether or not searching techniques should be transparent to the user remains problematic. The automatic implementation of the retrieval aids in Okapi'86 made it easier to test their effectiveness. In the Dewey project where the choice of options was left to the user and the search process was more interactive, results were less conclusive. This would depend not only on how search features are presented to the user at the interface but also on the type of retrieval aids being tested. It may be easier to evaluate retrieval techniques based on the matching or partial matching principle than contextual aids where there is an attempt to place the user query in a subject context.

The involvement of users at the different stages of the design process is an important part of the prototyping methodology. In addition the development of more interactive features does highlight the interdependence of user and system performance in the evaluative process.

2.2.3 Controlled experiments

A third approach in OPAC research has been concerned with investigating the cognitive elements of searching behaviour. Through controlled laboratory type of experiments and searching performance tests, these studies aim to gain a better understanding of how users search in order to ascertain how they can be assisted to search more effectively.

Based on the theory that the user builds a mental model of how a system operates, Borgman (1986a) investigated whether a user with a conceptual understanding of how a system works will perform better than a user with only a procedural knowledge. In comparing procedural and conceptual training (i.e. Boolean logic and operators), greater variance was found amongst individual subjects with different backgrounds (i.e. science/engineering vs. humanities/social sciences) than between experimental variables. Whether users with different characteristics require a different approach to training and how they can be trained to develop a 'correct' model of the system is unclear.

A second study looked at procedural problems for first time users, (Janosky, Smith & Hildreth, 1986). Subjects were given access to both online or printed help to undertake five assigned searches. Twenty four out of thirty searchers called up the online help in the course of the searches but success rates ranged from 0% to 58%. Some of the difficulties arose because searchers did not distinguish between procedural and conceptual instructions, e.g. type in 'author' was taken literally. Other problems were caused by not reading enough of the help screen which then made it difficult to recover from errors. The naive user undoubtedly with experimentation and a more

natural setting could overcome some of the procedural difficulties but the conceptual elements are proving to be much more problematic.

Nielsen, Baker and Sandore (1985) experimented with more in-depth instruction provided by a workshop and a printed brochure. Three groups of first year undergraduates were used, two provided with one type of instruction and a third, a control group received no instruction. All three were given written tests and online searches which were logged. Results showed that those who attended the workshop performed better on the written tests and marginally better on the online searches. However those who read the brochure did not perform better than the control group.

As with all exam type tests, questions can be raised as to whether the content or the method of instruction is being tested. In this case the learning objectives were very high and included selecting controlled vocabularies, truncation and Boolean logic. The authors concluded that some of the instruction would be more effective if it were 'embedded' in the system, (Baker, 1986).

Clearly user training and instruction in the use of online catalogues, cannot be regarded as a substitute for system improvement. However user assistance designed as an integral part of the searching process within the information system could make a substantial contribution to overall performance.

2.2.4 Non verbal data: transaction log analysis

With the advent of the online environment came the possibility of observing the user directly and unobtrusively without interrupting the search process. Automatic monitoring of activity on the computer system, i.e. logging transactions, was regarded as a powerful technique for evaluating user system interaction and performance. However the type of quantitative global data elements recorded, that is search commands, occurrence of errors, number of hits and time factors, did not provide sufficient information or the right type of data to inform adequately on how or why users searched in the way they did.

The method proved to have a number of more specific limitations. Firstly not all systems have logging facilities. If they do, specifications for the data recorded will vary from system to system, making comparisons difficult. Secondly there are problems in identifying individual search sessions as users do not log on or off as with other online bibliographic systems. Thirdly the log usually records users' input in full but the system's displayed output is logged in a coded form so that it is not possible to ascertain the basis of users' decisions. Fourthly logs generate a lot of data, the sheer volume makes analysis a daunting task.

If spite of these limitations studies of transaction logs have produced some useful diagnostic evidence of procedural and conceptual problems, (Borgman 1986b, Hancock-Beaulieu & Mitev 1989). Using direct observation to determine search boundaries, Borgman (1983) analysed search commands, types of errors, and time factors on the Ohio State University online catalogue. One in three of all sessions were multiple search types which included at least one subject search command. 13% of all commands were errors. Logs of the Melvyl system at the University of California collected in 1982 and in 1985, were compared and revealed changes in user behaviour, (Larson & Graham

1983, Larson, 1986). The call up of help screens had diminished. More searchers than previously were using the standard command search mode, as opposed to the menu search mode.

Another study analysed the effectiveness of keyword access in titles and searching LCSH headings. Failure analysis of transaction logs of three systems with different searching features revealed that searching LCSH or keywords in titles alone or a combination of both led to comparable failure rates (39% to 46%), (Kern-Simirenko, 1983). Searchers did not use available search options (Boolean, browsing headings or class numbers) to expand their searches but tended to change search terms instead. Two other log analysis studies by Dickson (1984) and by Henty (1986) revealed that most errors were due to miskeyings, mispellings or syntactic problems relating to initials, hyphens etc.

Transaction logging as a data gathering method has potential both as a diagnostic tool to be used in combination with other evaluative approaches in experimental settings or as a monitoring device for operational systems in general. The development of transactions logs is central to the work undertaken by City and being reported here.

2.2.5 Verbal data: protocol analysis

Protocol analysis has been used as a method of eliciting verbal data from users in two doctoral projects.

Dalrymple (1987) used a combination of observation and protocol analysis to compare how users reformulated a set of assigned searches in a card catalogue and in an online catalogue. The study found that the online catalogue seemed to stimulate more reformulations but the card catalogue led to the retrieval of more items. The author concluded that:

> It is not known how the internal cognitive reformulation process interfaces with the external system.

It appeared that the difficulty lay in isolating independent variables in the interaction and identifying the different types of feedback.

In an attempt to define characteristics of searching behaviour, Sullivan (1986) adopted a type of simulation method. A search planning exercise was designed in which six expert and six novice users were asked to plan 32 assigned searches. From the verbal protocols, search plans were produced which the experimenter then used to carry out the searches. The findings reveal significant differences in the planning process between experts and novices in terms of understanding the query, how they simulated the search and how they evaluate their likelihood of success. Novices concentrated on starting rules whereas experts focused on setting goals and planning alternatives in case of failure.

The work does point to some basic requirements for an 'expert' help system. However such a facility would need to be developed in parallel with improved features for a third generation system and not on the basis of current second generation catalogues.

2.2.6 Summary of evaluation methods applied to OPACs

It is evident that the methodological problems encountered in traditional catalogue use studies have not all been overcome in the online environment although some progress has been made.

The advances can be summarised as follows:

1. Transaction logs present the possibility of recording searches in their entirety.
2. Both system and user performance are starting to be taken into account.
3. Corroborative data is being collected in the form of verbal and non-verbal data.
4. A more systematic approach is being taken to isolate variables in testing new system features.

The state of the art in the evaluation of online catalogues is still at a diagnostic stage. It would appear that the development of more effective evaluation methods is dependent on being able to take full account of human factors at the interface.