Dilemmas of Document Communication

Axel Andersen

Royal School of Librarianship
The information explosion of the last century has had profound effects upon the entire society: The number of publications has increased enormously and still seems to be accelerating. At present something like 60,000 different journals of scientific, technical or other professional importance are being published, producing 3-5 million articles annually. In order to maintain a general knowledge of these immense amounts of documents it has been necessary to develop a modern library system and an improved cooperation between different libraries. In consequence libraries to-day are as much parts of an integrated network of institutions, as they are individual entities. Only by means of union catalogues and a narrow cooperation with a strong emphasis on the exchange of publications requested it has been possible to meet the challenges caused by the literature explosion and the parallel growth in number of requests.

Another important consequence of the rapidly growing number of documents has been the establishment and continuous development of new and improved bibliographic tools, starting with the foundation of the national bibliographies in the 19th century and the publication of subject and discipline oriented bibliographies in modern times - indexes, abstract journals and unconventional aids such as current contents journals and citation indexes. All these tools and the establishment of the international library infrastructure have been necessary conditions for the provision of relevant literature to users such as research workers and engineers in science and industry.

Computer technology has made a storage of bibliographic information in big, international data bases possible, and complex networks covering a broad field of subjects are now being established. Bibliographies, indexes and catalogues in their original paper-version are consequently disappearing and have, in fact, already disappeared.

Other sorts of literature are going to disappear too, being replaced by on-line accessible data bases. In France telephone directories are already disappearing in their traditional form now, and the same is soon
to be the case internationally with types of reference books containing big amounts of homogeneous data, such as address books, sales catalogues etc. The question is whether the "literature proper" - books and articles containing a coherent text - is going to be computerized as well. Already in 1969 F.W. Lancaster predicted that a complete change to a paperless society should be completed in 1990, and as late as 1979 Chris. Evans advocated exactly the same point-of-view.

Other new media such as microforms, video disks, teletext, and videotex are part of the ongoing diversification as far as information storage and transmission are concerned. At the EURIM 4 Conference, Brussels 1980, a half-way status was made. John W. Senders on that occasion reported of an experiment at producing an electronic journal without any paper version at all; this experiment, that was sponsored by the National Science Foundation, was not a success, but it may be a matter of time when the technological difficulties in producing such a journal are overcome.

It must be emphasized that there is a basic difference between two groups of new media:

On one side we have the on-line media in a broader sense of the word. They comprise all sorts of communication of information from some, common central store to individual users who have a potential access to the stored information, usually by means of telephone or television communication. Users are dependent upon access to the base of information, and this means of course considerable practical restrictions compared with information accessible in conventional printed form. On the other side this group of media have great advantages, especially with respect to their big capacity and the possibility of constantly keeping these huge amounts of information up-to-date.

The other main group of new media are of an entirely different type, consisting of prepacked, immutable units of essentially the same kind as the conventional media like books, other sorts of printed media, microfiche, gramophone records, music cassettes, and audio and video tapes as a whole. They are produced in a finished form which can in no way be changed, corrected or brought up-to-date.

Video disks seem to be an extremely promising new means of communication, especially due to its enormous capacity. One single video disk may contain 45,000 pages of text, or pictures - stills such as colour diapositives,
or shorter or longer sequences of pictures - for instance a whole motion picture with sound and music as well as pictures on the screen.

A Swedish government committee considering the possibilities of publishing a state-supported national encyclopedia in 1980 calculated that the content of a big, multi-volume publication of this sort can be packed in two video disks. The video disks are presumably being marketed in Europe within 2 or 3 years.

In the U.S. experiments are recently being performed, aimed at storing the text of entire books (The Bible, Huckleberry Finn) by means of microprocessors to be used in conjunction with a reading plate device on which the text can be blown up. Microprocessed books belong in principle to the same category of prepacked, finished products as their paper-version predecessors. They are still in their experimental phase but may very well have a considerable impact upon the entire communication market - and of course the graphic industries in particular - in the second half of this decade.

The accelerated expansion of the research sector has been accompanied by a corresponding growth of the information sector in general. Scientific and industrial results should be communicated to the international public as quickly as possible. Some consequences have already been mentioned: An increased publishing activity, the development of tools for a bibliographic control, and the establishment of a library system, all of them instrumental in the effective dispersal of information in document form.

As far as libraries are concerned the main task has not so much been the acquisition of new literature (which is certainly a tremendous task itself), as it has been the organization of optimal information retrieval systems, i.e. indexes, catalogues, bibliographies, and data bases, that might ensure the maximal and best possible retrieval of documents dealing with a given subject. In spite of all measures taken in order to improve professional communication, scientific society to-day is facing a tremendous problem of abundance.

Basically libraries always attempt at integrating the documents acquired in such a way that it is possible later on to retrieve all relevant documents. Classification is crucial in this context, and the establishment of most information retrieval systems involve classification, for instance in the production of a systematic catalogue, or a dictionary catalogue.
The big problem facing all librarians and indexers is the continuous obsolescence eroding all classification systems and systems of science. They are all reflecting the philosophy of their time, the Weltanschauung of the age in which they were constructed.

Comprehensive investigations have demonstrated, that the average, conventional library system assigns 1.4-1.5 subject references to each document. In modern computerized information retrieval systems a more exhaustive indexing is being performed, and the average number of subject entries per document (articles dominating) is roughly ten times as big as in conventional card catalogues; it means somewhere between 14 and 20 descriptors or other subject headings per document.

Comparative studies of a number of information retrieval systems unambiguously indicate that all the systems seem to suffer basic deficiencies, no matter if they are conventional or computerized. Apparently it is not possible to construct a perfect system: The Cranfield experiments show quite clear that the ordinary IR systems have a recall on 70-90 % of the relevant documents contained in the collection. Their precision is, correspondingly, 20-8 %; in a few, highly sophisticated systems precision is considerably higher. All this certainly cannot surprise experienced librarians. But on the other hand it has aroused considerable attention that the experiments indicate a clear correlation between recall and precision, to the effect that an improved recall is accompanied by a reduced precision - and vice versa. In other words: Improvement on one side is followed by loss on the other.

Ongoing research seems to confirm that there are limits to information retrieval performance. It is probably not possible to improve information retrieval beyond a certain point. But several factors affect this performance, and it is, therefore, difficult to establish those limits. Besides the quality of information retrieval systems, relevance assessment, and variations with respect to question formulation and reformulation, are important factors. And taking formulation of questions into consideration user studies may consequently contribute to an improved clarification of this important question. An analysis of the main types of search situations, and further investigations of the role of the human intermediary - librarian/documentalist - in the communication between documented information and human needs may also be useful in improving retrieval procedures by the aid of information research.
In almost any field, subject or discipline one can observe that the amount of literature is far too big to the reader. It simply cannot be absorbed, because it surpasses the needs, demands and capacities of the reader. The mass of documents relevant in the given connection is simply overwhelming. Consequently it is in most cases senseless to initiate a document retrieval in a given subject with the express purpose of procuring everything published on this subject. There simply exists too much.

For this reason search profiles have to be restricted as much as in any way possible. And even then we are very often facing an abundant literature so we are forced to put an upper limit to the number of entries we can accept - for practical reasons. This introduction of a purely quantitative upper limit is, in fact, just another way of giving up the maintenance of the bibliographic control.

The explosive growth of literature has created a problem of abundance which cannot be solved by means of data bases and computerization alone:

First the data bases usually give far too much information - too many references compared with the needs and capacities of the users.

Second the very lack of redundancy of the output is at the same time a strength and a weakness seen in relation to the actual search situations:

Apart from very specific subject enquiries performed by research workers, technicians and other experts, the great majority of typical document search activities are best characterized by the facts that the users do not at all want the entire mass of existing literature on a given subject. On the contrary: Users want to get one or a few, properly selected articles relevant in the given context. It is, consequently, highly important that user studies dealing with such badly defined search situations are being performed. They are, indeed, extremely common in public libraries with their mixed, and entirely heterogeneous clientele. User studies of this kind could therefore reasonably be initiated in this type of libraries (conf. T. Johansen, P. Ingwersen and P. Timmermann, 1980, and Annelise Mark Pejtersen, 1980).

Until recently user studies in information science have primarily focused upon users and needs within comparatively well-defined frameworks: Studies have almost exclusively been dealing with scientists, engineers, and other specialist groups, i.e. people with a common and rather homogeneous professional education and a strictly defined, usually rather narrow field of interest. Furthermore, investigations have as a whole concentrated on the communication of information by the aid of advanced, computer-
ized systems, and they have only sporadically dealt with conventional user-librarian dialogues.

Specific search situations of this kind can, moreover, be characterized by the large number of documents (especially articles) being retrieved, as well as by an explicit demand of the user to obtain an over-all, if not complete, view of the entire literature dealing with the subject in question. The persons in charge of this type of document communication are usually documentalists or research librarians with a university degree or some other equivalent theoretical background.

This advanced type of search - or rather: research - procedures are, however, not at all typical to the big and quantitatively entirely dominating number of questions arising from users in libraries of any kind. Students and "man-in-the-street"-borrowers are in several ways in extreme contrast to the research workers with narrow, advanced, and well-defined questions.

The extremely rapid development that has taken place in the field of information technology and information retrieval systems makes this bias quite understandable. It is, however, highly important - for the advancement of information science as well as for the improvement of the library practise as far as service to readers is concerned - that new approaches to user studies are being introduced in order to obtain a broader understanding of the basic factors 'use' and 'needs' than we can possibly get by means of the mentioned, well-established investigations in user studies dealing with a very small and special clientele. It is necessary to expand the field of research, especially with contributions which can shed light upon the huge majority of rather simple and for several reasons badly structured search situations. They are, indeed, most predominant in public libraries, especially in the smaller units, but their rôle is by no means unimportant in research libraries and in libraries serving institutions for higher education. As seen in contrast to the rather homogeneous users in the group of specialists, public libraries as a whole are characterized in serving an extremely heterogeneous group of readers - children, young and old people, persons of all sorts of occupations, students, handicapped, immigrants, unemployed, uneducated, etc. etc. Most striking is maybe the enormous differences as far as their educational background is concerned. But the search situation is of course very much influenced by the very size of the document collection at hand in the library.
Some attempts have recently been made at attacking the user studies from this angle - the public library aspect so to say. This goes for subject-oriented searches as well as inquiries concerning fiction - reading of novels etc. It is certainly not surprising that these badly structured user interests have been left untouched until recently. They seem to be more difficult to analyse than search procedures in fields of science having a well-established, fully fledged terminology and system of concepts.

Even if this distinction between advanced, professional users on one side, and the non-professional users of the public libraries on the other, even if it is obvious and to some degree justified, it may nevertheless be misleading: probably we should rather choose to consider the in-depth search as exceptional, and the more "superficial" (and consequently highly selective) approach to literature as the predominant, all-important type. They are so, not alone so far as public libraries are concerned but also for the other, more specialized document-communicating institutions.

The study of communication processes in information retrieval and in libraries can be a direct benefit to teaching and training. This includes the training of librarians and information officers who are mediators, and the training of users. User training is often performed by libraries without any coordination with other study activities. Training in library use should be closely connected to subject training, because the subject knowledge in terms of terminology and overview of the field is a prerequisite for meaningful search procedures. In the educational field there is a well established connection from teaching practice over teachers' training college to educational psychology and further to psychology and sociology.

In library training and education in information retrieval there is a similar link through information science to linguistics and sociology. But the link is weaker and information science has mainly dealt with problems concerning communication of scientific and technological material and retrieval tools with emphasis on computer retrieval, automatic indexing etc. Research on communication processes in the humanities and not least in the public library sector is far more scarce.

Traditionally there has been a link between information science and linguistics. We have found that theories in cognitive psychology and psycho-
linguistics may explain many phenomena in search processes and communication processes. There is also a mutual influence between linguistics and psychology on one hand and computer science and artificial intelligence on the other. The parts of artificial intelligence dealing with question-answering systems, automatic translation and problem solving may develop in a way that could have a profound influence on information retrieval and reference work.

In recent years a cognitive viewpoint or cognitive science has developed in the different scientific disciplines as a unifying force. In educational psychology this has been advocated by Ausubel and Novak. There is good reason for exploring this trend in information science and library work.

References:


FEJTERSEN, Annelise Mark. Investigation of search strategies in fiction based on an analysis of 134 user-librarian conversations. IRFIS 3 (3rd International Research Forum in Information Science). Copenhagen,
