PP-76-8

USES AND ABUSES OF DATA BANKS

Peter Popper

September 1976

Uses and Abuses of Data Banks.

Peter Popper

Lecture to Polish Academy of Sciences, Warsaw, Oct. 1976.

Abstract.

After reviewing the historical development of libraries and catalogues - indexes -, the evolution of secondary or abstracting services is discussed in the light of the rapidly growing rate of publication of new research literature. The retention of traditional formats and indexing practices is shown to have had an adverse effect on the benefits computerization of abstracting services might have brought. Users of such services have to contend with a multiplicity of systems which makes maximum utilization difficult. Inspite of the large volume of abstracts published, evidence is presented to indicate that only a relatively small proportion of literature is currently reported in secondary services. Rejecting a possible model for a global information system, it is argued that addition of citation indexing to secondary services would offer a means of screening the literature for the "valuable" publications. After discussing the utilization of literature in increasing the knowledge domain of users and the influence of user habits in using knowledge in decision-making, possible future developments in information systems are outlined.

It is my intention today to discuss some aspects of modern data banks, which perhaps have not always received the attention they deserve, and to show how these features are to some extent the product of the historical evolution of data banks. I then want to deal with what to my mind are the shortcomings of existing documentation systems. It is, however, not sufficient to point out deficiencies, and so I will try to suggest solutions to the problems posed, discussing existing proposals critically. Naturally, there is much that is sound and good, and these aspects will be explained and amplified. I also wish to speculate on what the features of data banks may be in the near future, say the next fifteen years, and in the further future, say fifty years, bearing in mind the probable progress in ancillary technologies.

I should start by putting those of you who are not computer experts at ease: neither am I. I, like most of you, am a mere user, who perhaps has taken a more critical look than most at what is being done as against what could or should be done than most, but then I have been in the "business" for a long time. I would also like to define some of the terms I shall be using. Data bank, unless specifically otherwise indicated, will apply to a bibliographic data bank, a library-type system, in which each entry for a single bibliographic entity will be termed a record: this is made up of various fields and contains information which will be termed the "bit" of information.

We should also look at what the "information business" is. Basically it extends from the producer of new information, the inventor, innovator or writer, to the user of the information, who in turn probably also is a producer. We are therefore concerned with the transfer of the bit, whether this involves transmittal by word of mouth, the written word or the computer record. It is both a closed and an open system with many loops and many paths leading from A to B - a network if you want. We in the library and information business are merely involved in ensuring speedy transfer from A to B, in storing the bits so that present and future users can get at the information. Everything we do must keep these salient facts firmly in focus. The transmittal system must never become the end in itself, but must always remain the tool.

It is sad that after all the years over which attempts have been made to bridge the gap between originator and user, all evidence points to a dismal failure of past and present systems. For if we look at how information users acquire their information, we of the "information business" stand condemned. How does a researcher acquire wanted information, how does he keep himself informed: the sources are original articles, books and reports; secondary sources souch as abstract journals, library catalogues or data banks; and direct communication, oral or written, with and from colleagues. Many surveys have been made of this "information" intake. Whilst all produce different figures, the variations are slight compared to the major breakdown: more than fifty percent comes from direct sources and less than ten percent from secondary ones.

These are damming figures and we really ought to see why this is so. This too is an aspect I will want to touch on later. Let me state here that I believe that there is no single reason which applies to all situations, but a multiplicity ranging from "unfriendly" systems to work in some new direction for which no system exists.

Libraries have a very long history going back to Assyrian times or even earlier. Probably the first collection which could call itself a national or royal library was the archive of cuneiform texts assembled by Ashurbanipal at Nineveh about 650 b.c. With the recognition that the written word exercised some sort of magic power over tribes, it soon became common practice to place the safekeeping of such archives in the hands of priests; and so temples, subsequently monasteries, became the repository of manuscripts and later on of books. Indeed there are no monasteries or similar institutions throughout the world today that do not boast of an extensive library housed in the best and finest of the rooms available.

None of the monastic libraries, of which that at Monte Cassino dating from 529 a.d. is taken to be the first, was large by modern standards. Furthermore, although the earliest archives happened to be collections of legal and administrative, sometimes financial, documents, monastic libraries tended with the passage of time to become more and more ecclesiastical. It was only the coming of universities which changed this trend, although even then it was priests or monks who acted as keepers or librarians of the various collections. With the revival of interest in Latin and Greek texts and the spread of humanism there followed a new interest in book collecting, both as an aid to scholarship and as an end in itself. Petrarch and Boccaccio set the pattern for the working library for scholars, and wealthy patrons became book collectors with possession of a library being a symbol of wealth and status.

With the evolution of libraries, it was not long before librarianship itself became an object of study with the first such book, "Avis pour dresser une Bibliotheque" by Gabriel Nude, the librarian of Cardinal Mazarin, appearing in 1627.

The scientific journal did not make its appearance until the seventeenth century, the first two being the "Journal des Scavants" and "The Philosophical Transactions of the Royal Society". Subsequent growth was rapid: by 1750 there were ten serials, by 1800 the number had increased to 100, there were 1000 in 1850, 10,000 by the turn of the century, and probably more than 100,000 today - a growth rate of one order of magnitude every fifty years.

It was therefore not surprising that the problem of how to gain access, how to find a given record or the contained bit, became a focus of interest. Even the earliest libraries, certainly the papyri collections at Alexandria, kept records of their holdings: the beginnings of catalogues. nor was it surprising that an important part of a catalogue entry concerned itself with the description of the manuscript or book: these were rare or even unique and had to be clearly described as a means of identification should they ever become misplaced, or more probably looted from their location. It was also not really important to ensure that descriptions or other cataloguing characteristics were uniform and standard: with small collections, a user could soon learn a new system on moving from one library to another. With the increase in volume of material housed in libraries, and more importantly the overlap in holdings among libraries as printed books became available, uniform cataloguing or indexing became of greater significance. The pioneer work in classification was done by a sixteenth-century Swiss doctor, Konrad von Gessner, and a group of French booksellers in the seventeenth century. Melvil Dewey produced his decimal classification scheme for Amherst College in 1872 and this system was rapidly adopted by other libraries, and information systems, particularly in the English-speaking world. Partly in response to a demand for a quick agreed development in rapidly expanding branches of knowledge and for classifying highly specialized material, the Universal Decimal Classification, UDC, was evolved from the Dewey system. It is currently available in a large number of languages and is widely used throughout the world. The U.S. Library of Congress also published in 1902 the schedules of its own classification system, which too found wide adoption.

The period to the mid-fifties of this century basically saw merely a refinement, and expansion, of the old-established techniques.

About a century earlier, when there were still fewer than a 1000 journals, it had already become apparent that researchers could not keep adequate track of developments in their fields, and some journals, such as the "Journal of Chemistry", started to include in their pages short summaries of papers published elsewhere - the beginnings of abstracting or secondary services. With time, these services became separate entities, and like the reason for their existence - many serial titles - they too started to multiply leading slowly to the establishment of a virtual industry - the "information business". There were several early attempt to bring order into the threatening chaos; thus The Royal Society published up to the turn of the century a bibliography, claimed to be comprehensive, of all articles appearing in serials, but had to give up the attempt because of the size of the task involved.

In parallel, there was the growth in patent literature with its attendant problems of classification - naturally different for each country - leading to problems of establishing priorities etc. Possibly here the Europatent, and ultimately the Worldpatent, may bring some alleviation.

What we thus had at the beginning of computerization was the set habits of libraries in cataloguing their holdings, and the various abstracting services, each working to its own system. Computers were brought into the information and library business in the hope that they would provide the universal solution. Why, it was argued, should not machines perform the thankless task of sorting and arranging millions of catalogue cards which in the case of abstracting services had to be typeset, proofread and all the other wearisome stages of publications.

Unfortunately, tradition was difficult to overcome: much effort and ingenuity went into solving the problem to how to make the machine produce a mirror-image of what had existed before. No real thought was given for quite a long time of how the capabilities of computers could be fully exploited to the best advantage; on the contrary, steps were taken to make systems inefficient to preserve the outward appearance of the "steam"-produced services. Thus "Chemical Abstracts" even today uses some twenty-five typefaces, with all that involves in redundant control characters; records still contain information on the size and nature of binding of books - merely because tradition has so decreed. Is an information user really interested in knowing that the book he ought to read is leather-bound and 25 cm. x 15 cm. Personally I doubt it.

The same approach also applied to indexing. The UDC system certainly is excellent, but why should an information user first have to look up an extensive and complicated code book to find his relevant set of numbers, when machines allow a free text approach, and even can add structured thesaurus facilities, the raison d'etre of the UDC? True, a code classification can be helpful and space saving, even in a computer system, but it must be relatively simple and easy to memorise. (We at IIASA actually use a coded broad classification scheme, but there are only some 50 classes of a single letter/single numeral type, and this never appears on a printed output in order not to confuse the user.)

The driving thoughts behind the introduction of computers to libraries were cost savings, speed of information dissemination and completeness. Let us see how we have failed to achieve these aims. The pioneers in the early efforts of using computers for information handling were the American Chemical Society - "Chemical Abstracts", the American Society of Metals - "Metal Abstracts", and the American Library of Medicine - "Medlars". The Library of Congress too launched its computerization program, leading to Marc (machine-readable catalogue) tapes. Each development was expensive, swallowing and continuing to absorb large amounts of money in design and implementation, and now in refinement and modification. Each of the systems has many merits, but each is different and totally incompatible with others. Indeed even today new systems are being designed and established with little thought of aiming at any degree of compatibility.

Certainly of late there have been attempts to bring some form of standardization into systems and the records they keep. But even here there is a multiplicity of international standards and recommendations - Purim, Iso, Unisist - each carefully being different and clinging to the historical roots: different bibliographic fields for different types of publication, different methods of sequencing for bibliographic entries etc.

Cost saving is the least worry of systems designers and no thought is given to the costs imposed on the user in having to deal with the plethora of formats.

As regards speed of dissemination, the large computerized secondary services certainly have nothing to be proud of: if one remembers that for instance the prewar manually produced "Chemisches Zentralblatt" had an average lag between apperance of an original publication and notification in the abstract journal of about eight weeks, then the currently accepted delays of three to four months, and very much longer in many instances, are certainly no progress. Indeed it was to some extent this growing delay which led to the introduction of Selective Dissemination profiles - there were and are other very cogent reasons why predistribution of selected portions from the master tape became necessary: part of the time lag in producing the final "printed" version is indeed due to the variety of typefaces used - the relict of steam-produced versions.

Finally, there is the question of completeness. Anderla, in his study "Information in 1985" has published figures for the number of abstracts reported in nineteen major western secondary services. For 1957 there were 600,000 and for 1971 1,000,000: extrapolating to 1976 gives a figure of about 1,750,000. An optimistic figure for the remaining services which are now estimated to exceed 2000 (and some of these publish as few as 200 - 500 abstracts per year) gives about 5,000,000 abstracts. Since many original publications are reported in several services we must make allowance for such duplication: a fair estimate is that each original is reported at least twice, so that the figure must be reduced to some 2,500,000 separate bits of information.

Against that we must put the number of original bits published. I have already mentioned a figure of 100,000 journals currently appearing. Each of these will contain on average some 100 papers a year, giving a figure of 10,000,000. To this must be added the reports, patents, books, conference papers etc. These will probably amount to the same number, giving a grand total of 20,000,000. Compared to the 2,500,000 bits reported, the figures are certainly disturbing.

There is other evidence to support the above findings: Maurice Line recently published some highly significant figures relating to usage of the British Library, Lending Division, at Boston Spa. This currently holds nearly 50,000 serial titles, and during 1975 supplied close to 2,500,000 photocopies of articles. Yet a survey made showed that less than 15,000 serial titles satisifed all the requests made during the three-month survey period, and of these 3,800 were no longer current, i.e. had ceased to appear. 80 percent of requests were met from 5,200 serials, 50 percent from 1,400, 30 percent from 450 and 10 percent from 58 !

Discussing these figures with Maurice Line left both of us somewhat puzzled; this all the more so in that the most commonly requested titles were actually journals one would expect to find in any specialist library in the field: the list of these was headed by "Science", followed by such titles as "Nature", "Jnl. Amer. Chem. Soc.", "New England Jnl. of Medicine", "Jnl. of Biological Chemistry" etc. There are several possible reasons for these remarkable figures:

1. The unwanted titles are of poor quality. This is disproved by the high-grade titles appearing in the unused list.
2. Availability: the high-frequency-use journals are all held by more than thirty libraries in the U.K. according to the most recent issues of the British Union Catalogue of Periodicals, which covers only a limited number of libraries, and certainly does not include industrial libraries.
3. Language of journal: some thirty percent of Boston Spa usage is now from outside the U.K., so whilst there may be a weighting in favour of English language titles, there should certainly be no exclusion of non-English ones. In any case the unwanted list contains as many English as foreign language titles.
4. Specialist nature of the journals with a very small readership, with most or all specialists having their own copy. This may account for a small portion, but is certainly not the main reason, especially considering the high cost of some of the specialist titles.

Considering that some of the unwanted titles are actually the only copy held in the U.K., one must look for other more cogent reasons for the strange usage figures. Dare one suggest that possibly the very titles of the serials, let alone of the articles contained therein, are unknown to most? And if this is true of serial literature, which was and is the easiest to control from an information handling point of view, what then about report or conference paper literature? It is thus quite obvious that the secondary services must be failing in their aim of being comprehensive.

It is perhaps surprising that to my knowledge no real studies have been made of this aspect. A fairly easy and cheap method would be to use a large library such as Boston Spa, which claims to receive every secondary service published worldwide. If one took one punched card for each article in every serial received and encoded the abbreviated essential information, ISSN (International Standard Serial Number), year, volume, part number, first page number and first six characters of the first author's name, together with date of receipt, and did this also for every entry in every secondary service received, it would be simple to check whether an article was included in a secondary service, what degree of duplication obtains, and the time delay involved in notification. (Presumably outside time lags due to postal handling etc. would be equal for primary and secondary publications).

I suspect that such a study would disclose some highly unpalatable facts: excessive duplication, unwarranted delays, and a startling number of non-notifications. Possibly even such nationally comprehensive services as Bulletin Signaletique in France and Referatny Zhurnal in the USSR might be surprised at the findings.

What steps could be taken to overcome these deficiencies? Cost savings - here simplification would go a long way, but greater benefits would come from standardization, making services compatible and data exchange possible. Both these factors would also contribute to reducing time delays. As regards comprehensiveness or completeness, that is a much more complex problem and really also involves the question of whether this is ultimately desirable from a user point of view. But let us for the time being assume that we are aiming for completeness.

The outlines of how to achieve this already exist: Unisist has proposed, and some large systems such as INIS, AGRIS, already have adopted, regionalized input. The difference here would be that input from any one centre would not be subject-oriented, but total for the geographic area concerned; i.e. Poland would have a centre responsible for preparing input to a global centre for every journal article, report, book etc. published in Poland irrespective of the subject matter. The initial input should comprise the essentials, that is title in original language, title translated into the system language or languages, authors, affiliation or corporate authors, standard bibliographic reference, possibly an abstract, and very rough keyword or subject codes - these being to a strict pattern and confined to the minimum. Input should ideally be in machine-readable form. Additionally there would be an identifying record number. The entry - and this is the important point - accompanied by a copy of the original would be transmitted to the global centre, sorted by subject codes, and sent on, again with copy of the original, to one or more subcentres for more detailed indexing, and if subsubcentres are required, on to these - chemistry, medicine, physics may well be cases in point. There could be several levels of subcentres, each enriching the indexing as required. It might even be feasible to feed back the indexing enrichments to the initial entry at the global centre. We would thus have a comprehensive system, which would guarantee totality of input.

A user would therefore be certain that any information request would be met by a total response. Furthermore, the global system and its subsets would be compatible and to a standard, with the further assurance that each sub- or subsubcentre would also have a copy of the original material. I have not concerned myself here with the questions of precision and recall, assuming that the indexing is perfect. (We know that this is not so.)

But would we as users really wish to have all the literature for a given topic? Could we actually absorb and process this for our particular needs? Could we handle a "Chemical Abstracts" which might be ten times its present size, would we really want "Medlars" to give us 3000 case-histories instead of the present 300, for some obscure syndrome? I would suggest that what we really want is the best 300, but who is to assess the [illegible] to make the necessary selection? Garbage in, garbage out has long been the most valuable adage of data systems: what is more, garbage breeds garbage. Furthermore it is frequently true, that what is the garbage put out by one system is the "valuable" intake of another. As it is, we are already suffocating with too much poor information under the present regime.

Quite apart from any other considerations, such a global system, with subcentres far apart, would inevitably result in even greater time delays than at present, with virtually no offsetting advantage.

There is a further shortcoming of the present state of affairs: the subject scope of existing data banks shows severe gaps. There are very few "soft" science data banks, for fairly obvious reasons. Historically, most large data banks were backed in their development by some commercial interest, i.e. the chemical industry needed "Chemical Abstracts", the pharmaceutical industry "Medlars" and "Excerpta Medica", etc. There was no similar support for the soft sciences which are not normally industry based. It is true that currently with the appearance of many new paradigms small manual services grow up in those areas where there are no clear-cut "core" journals: the normal path is from newsletter to specialist journal, or where that fails to become accepted, to a small secondary service. So these areas at present tend to be excluded from automated data banks, although again there are various attempts such as Devsis or Spines.

We have at our disposal however a valuable tool which could be applied as a yardstick to measure the worth of an original publication: subsequent citation. Much work has been done on the value and use of citation analysis, and the "Science" and "Social Science Citation Indexes" are amongst the most important bibliographic search tools available. If we postulate that a "valuable" paper will be subsequently cited by other researchers, then we could say that five citations, other than self-citations, are a measure of quality. It is true that at present not all published papers are universally known, but with a global system this deficiency would no longer exist: following initial notification, entries in global data banks would be held in an interim form and would only be included in the main reference data banks if they fulfilled the criterion of being cited five times over a given period. Reviews of what is included in the reference data banks could be made periodically to ensure publications whose worth is recognized only many years after initial appearance in print being transferred to these data banks. Indeed, Garfield already uses citation frequency of journal titles to expand his present data bases.

We must however be certain that our analysis is correct. A large system is far less likely to function than smaller ones; the people concerned with it are too far removed from the point of impact on users and will tend to have little interest in whether the system really works or not. After all, any system has an effect on the people within it. It isolates them, feeds them a distorted and partial version of the real world outside and gives them the illusion of power and effectiveness. But systems and people are related in another subtler way: a selective process goes on whereby systems attract and keep people whose attributes are such as to make them adapt to life in the system. And in dealing with a precious commodity and resource such as information, attracting and keeping people as operators who are not interested in what the system is supposed to do, that is transfer information efficiently and quickly from those who have or are producing it to those who use it, is not going to give the results we want and need.

There are certain aspects of any system which apply particularly clearly to information systems: big systems either work on their own or they don't, and if they don't nothing will make them function - pushing on the system will not help either. On the other hand, if a system works, it should be left alone. Complex systems that work have invariably been evolved from simple ones that worked. Indeed, a complex system designed from scratch is unlikely to work. It is therefore unlikely that a global information system as outlined above would work and satsify the users, all the more so since a large system produced by expanding the dimensions of a smaller working system will not behave like the smaller system, from which it has evolved.

It is therefore almost foolhardy to suggest that merely expanding one of the existing large data bank systems would give us the required results. We must really look further and include in our analysis both the producer and user of information, and not only concern ourself with the central transmittal portion.

As regards producers, we could reduce the size of the problem materially by eliminating duplicate publication. At present we all tend to have "writitis" if one may coin a new expression. Because of the pressures of promotion races where number of publications is considered of paramount importance, we all tend to produce several versions of any new bit of information for publication. There was this summer a seminar in the U.K. where this view was roundly condemned and a recommendation made to British universities to change their approach. The plea was made by users of information and the publishing world, and backed by a large section of academia. It is probably reasonable to estimate that each new finding is reported at least three times, and reducing the total number of publications by this factor or more would give the single largest relief to the problem of information handling.

Many producers of information are themselves acutely aware of the shortcomings of the information dissemination business, and have taken steps to shortcircuit existing systems by setting up their own private distribution networks, the so-called "invisible colleges", where the fifty or hundred researchers in a given small area communicate their results and findings directly amongst themselves. Newcomers can easily enter these "invisible colleges" by indicating their interest in the research publications of one or two members of such a community - they will be welcomed and absorbed. The one thing which must never be done is to attempt to formalise these networks, for this will destroy them without fail.

The producer here is also the user and the distributor, and since the system is efficient, it should be left alone. Expanding it to cover more than one topic for a given network will destroy its very reason for existence. Many researchers are actually part of several such networks, but there are no known successful syntheses of networks.

We are all also users of information from birth till death. Indeed information, the state of knowing, is fundamental to life itself. Deny the organism the process of acquisition of information and death will ensue. But information is also a commodity, a matter of worth, of acquisition and of power. We should however also remember that data bases are only a peripheral part of the knowledge domain. When someone is informed of some knowledge or data he will transform this to information and be in a different state than before he was exposed to the knowledge. So the response function of an individual is a fundamental part of the knowledge domain. The satisfaction of the need for data or knowledge can be considered to have intensity, directionality and all the attributes of other human satisfaction. Translated into our context this means that we as scientists and technologists have a need for the food and water of knowledge, which finds its consummation in the process of taking data and knowledge from the event world - the published and verbal communications - and transforming them into useful information.

This concept has underlying implications.

First, each of us satisfies needs in specific ways; there are individual differences in the way we express and satisfy our needs, and this applies also to information. Thus some prefer the spoken to the written word.

Second, the satisfaction of need is a social as well as biological affair. The way we attend to our needs is strongly influenced by the way others around us attend to their needs. For example, if we see everyone at breakfast in a hotel reading a special issue of a newspaper, we too will wish to follow suit.

Third, needs are satisfied through the organized mechanisms (the systems) that are available to us. They are molded by the environment in which they find expression. Again, if the society in which we live produces a morning paper, we will read this. If newspapers appear only in the afternoon, our habit will change.

Fourth, needs are modified by education and training. We learn to moderate our requirements, expressing them differently and integrating them with other needs. To use the example of newspapers yet again: if we have available twenty daily papers, we will soon learn which one or two give us the information in the way we can best absorb it and which best meets our selection criteria.

Information represents the end state of a process initiated as the result of a basic need. There are some limiting factors to be considered.

First, awareness of acquisition: the ability to know when to acquire and where to acquire.
Second, capacity or limits of acquisition. Given an array of objects that are related to our need, what is the capacity of acquiring these?
Third, deterrents to acquisition. When presented with an array of objects or facts, some of which are related and some of which are not, what are our limits in sorting and acquisition?

There is also the interesting fact that the more you are exposed to knowledge, i.e. books, articles, etc., the less alert you become to knowledge in general. However, the threshold of awareness for things you are interested in is lower than your general threshold for new knowledge, and this offers some alleviation. It should also be remembered that the overall storage capacity of the human brain is somewhere of the order of eight to fifteen powers of ten expressed in conventional computer bits, that is, greater than the memory capacity of the most advanced computer. Actually, human capacity is even greater when we allow for the combinatorial powers of the human brain. Limitations arise not in storage, but in retrieval of information.

We thus come to the question of utilization or matching information resources to human abilities. Data or knowledge become information when we use them, and in turn, information is data of use in decision-making.

Again there are limitations in this. Decision-making and problem-solving are directly linked to cognitive style and to personality. There are several relevant aspects we must consider:

1. The amount of data we need to support our actions varies with our perception of the world or the problem involved. Those who are conservative in their estimation of events require more data than those who are risk-oriented. Both demand more data or knowledge than required, and when this is provided, it is seldom used.

2. The way in which we solve problems and make decisions depends largely on our attitude to the problem.

3. We normally tend to solve problems by a step-by-step approach.

4. Individuals adopt different strategies in their attempts to understand the attributes of information available to them.

5. Decision theory suggests that the way we act on alternatives is to postulate the probabilities of possible occurrences; the probabilities in turn are influenced by prior

experience, both actual and subjective.

All this leads to the sad conclusion that the hope of a single design, a unique algorithm for the setting up of information systems is a myth, an impractical dream, which should not be pursued. If there is any criterion for data bases or information systems it is that they should be as flexible as factors related to their operation will allow. We are dealing with a dynamic system: and it can only function well if it allows these dynamic processes to proceed correctly for the task and environment involved.

Using a given data bank will often not give a result relevant to the user's thinking style. Data banks are usually structured on a logic that is not that of the user. On-line data base usage can reduce the effects of the disparity, but it will require the user to adapt to the logic of the data bank structure. And the real nub of the problem is that each data bank has its own structural logic.

Our short term aim should therefore be to make at least the logic of data base structures compatible. It is patent nonsense to extend the demand for compatibility to include indexing terms or thesaurus structures. Nor is it meaningful to complain if items appear in several data banks - they may rightly belong there. We should complain of the absence of relevant or valuable bits and the inclusion of trivial material which only increases the noise level. We should further see that such restructuring as is done should be based on the best of the present systems, and that data banks are set up for those subject areas at present covered indadequately only or not at all. covered. Far better to have even 100 good working data banks for 100 subject areas than one global one which will not work at all.

The utopian dream is to have information available on the day of publication at the latest, neatly translated into one's mother tongue and packaged in collections of bits which are of infinitely variable size and content. Whilst the ultimate realization of this dream is probably not feasible, there are some things that can be done to translate it into a workable program, but these will involve radical changes in the whole concept of information handling.

We must start thinking in terms of information as a physical quantity, especially as regards library materials. With access to information being at the user's desk through terminals, we can no longer retain our traditional library: this really becomes a warehouse with its standard material-handling problems. Instead of nuts and bolts of a given size or metal, we are now dealing with books or journal articles. Indeed, this has already been realized in Boston Spa, for example, and other libraries or document-delivery services are following suit.

Over the next fifteen years I would hazard that we will see libraries divide into two groups: small in-house specialist collections, frequently with contents varying with research interest, and large regional document-delivery systems. In the further future, document transmission may achieve the necessary technological breakthrough to allow on-line usage in a system parallel to computer networks. Possibly the two may be combined, but I do not believe that full text computer storage of all written matter is a practical consideration.

Networked data bases are already a reality, and the near future will see an ever growing expansion of networks and available data bases. However no data base without adequate document-delivery services will be able to survive. First-line material, i.e. that in reference data banks, will have to be available to the user within three to five working days, secondary material can suffer longer delays. Indirectly this will also have an effect on the publishing industry: document-delivery centers will become prime customers, to be cherished and supported, and not, as at present, considered a dire threat to an industry now fighting its own self-made difficulties.

In my view, inclusion in present data banks of citation indexes is also a certainty; the time horizon is questionable and will depend to some extent on developments in computer technology and network facilities. If data bases exist, then access to these must be free and rapid. The present difficulties in networking due to inadequate line facilities or expensive line charges will have to be solved. After all, we spend quite a significant amount of gross national product on research, and then make access to the information obtained prohibitively expensive and difficult. In the longer term I believe that some of the data bank input may be provided by the primary publisher - either, where using, computer type-setting by passing on portions of relevant input to secondary services, or by including some pages containing the relevant information in an agreed format amenable to optical character-recognition processing.

We will also see over the next few years a growth of referral centers, which will be the first point of call for information requests. It will be the function of these to tell users which data bank is likely to contain the relevant information, a sort of super-index. Again, there will be many instances, especially in new disciplines, where more than one data bank will have to be approached to search exhaustively the literature. In the long term again, such referral centers may actually turn into data banks themselves. There are some beginnings here already - the Environment Referral System of U.N.E.P., which will be online, soon may be a prototype. Such a referral data bank may possibly use citation analysis to point users to actual researchers or institutes working in a pertinent field, thus bypassing the literature altogether. In this way data banks could support rather than destroy "invisible colleges". There will also be a growing trend to make actual numerical or physical data available: many statistical or physical data banks already exist. Indeed there were attempts some years ago by the German steel industry to build up an interactive properties data bank for steels which would allow the selection of a given grade of steel from the properties required for a particular application. Another examples is the Mayo Clinic medical records system which allows physicians to use earlier case histories in diagnosis and treatment.

As regards the problem of language difficulties, here too much progress has already been made and this will accelerate. I.B.M. over ten years ago established its own in-house system which is multilingual, with the machine translating keywords. The Titus system of the textile industry is another example, and the Road Research system will have similar facilities. There is no reason why multilingual facilities should not be included in other data banks. Hierarchical thesauri of index terms with automatic upward posting have also been incorporated in some systems, and usage will extend. Whether automatic text analysis and indexing will become commonplace is slightly more problematic: if synopses are written to a strict regime, then they could be used for automatic inverted file construction, but this will require much discipline on the part of synopsis writers. We still know too little of the semantics of free text writing in all languages to predict a definite future in information handling, especially as full text storage - a prerequisite - is also more than questionable.

Whilst there are certainly many deficiencies in our present information systems or data banks, I do not view the future as being gloomy. On the contrary, much is being done to improve the situation and progress is relatively rapid. As long as we realise that we have a real function to perform in the "information business", that we are dealing with real users, we shall make progress. If however we get lost in utopian dreams where we consider the system per se more important than what it is supposed to be doing, there are real dangers ahead. Small, or relatively small, may well be beautiful.