DOCUMENT RESUME

ED 127 802 FL 007 922

AUTHOR	Ebersole, Joseph L.
TITLE	Some Probable Technological Trends and Their Impact on an Information Network System. LINCS Project Document Series.
INSTITUTION	Center for Applied Linguistics, Washington, D.C. Language Information Network and Clearinghouse System.
SPONS AGENCY	National Science Foundation, Washington, D.C.
REPORT NO	LINCS-3-70
PUB DATE	May 70
GRANT	NSF-GN-771
NOTE	17p.
AVAILABLE FROM	National Technical Information Service, Springfield, Virginia 22161 (PB 192 494, MF $0.95, HC $3.00)
EDRS PRICE	MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS	Automation; Clearinghouses; Computer Output Microfilm; Computer Programs; Computer Storage Devices; Indexing; *Information Networks; Information Processing; Information Retrieval; *Information Science; Information Services; Information Storage; *Information Systems; *Microforms; Program Costs; Resource Centers; *Technological Advancement; Time Sharing

ABSTRACT

Improvements in the technology associated with the information sciences will have their primary potential impact on the distribution of costs, information flow level, information availability, and use among information channels. This improvement implied not only a capability to perform a given function, but a lower cost. For example, the trend towards multi-access computers implies cheaper and more accessible processing. In considering storage costs, microfilm appears the most economically feasible for new information systems, although this presents new problems regarding the number of data banks and accessibility. The expansion of the micropublishing field should result in reduced relative costs, which in turn would mean greater availability. Vertical integration of channels suggest some answers for active subject areas with a high publishing rate and sizable community. The development of low-cost mini-computers has made possible searches by a remote user. Finally, the creation of standard formats and transferable computer programs will allow for tapping other data banks. (Author/AM)

CENTER FOR APPLIED LINGUISTICS

LANGUAGE INFORMATION NETWORK AND CLEARINGHOUSE SYSTEM (LINCS)

SOME PROBABLE TECHNOLOGICAL TRENDS AND THEIR IMPACT ON AN INFORMATION NETWORK SYSTEM

By

Joseph L. Ebersole

LINCS PROJECT DOCUMENT SERIES / NATIONAL SCIENCE FOUNDATION GRANT

LINCS #3-70 May 1970 NSF GN-771

CENTER FOR APPLIED LINGUISTICS, 1717 MASSACHUSETTS AVENUE, N.W., WASHINGTON, D.C. 20036

SOME PROBABLE TECHNOLOGICAL TRENDS AND THEIR IMPACT ON AN INFORMATION NETWORK SYSTEM

By

Joseph L. Ebersole

CONTENTS

1. Introduction

Statements which predict technological advances and estimate the time period for their occurrence normally contain an implicit value judgement that what can be done should and will be done. Many technological advances are, in fact, applied by decision makers who seem (especially with hindsight) to have tunnel vision and who relish the euphoria of the ever accelerating trip through their ever extending tunnel. Simply stated, we should not do something merely because it is possible to do it. Therefore, the emphasis here will be more on the implications of achieving the capability to do new things (or to do things which were not heretofore economically or technically feasible), with the realization that when a given capability becomes available, this in itself should not be the sole determinant of systems planning. These considerations are presented in the light of certain future requirements of the Language Information Network and Clearinghouse System (LINCS) now being planned by the Center for Applied Linguistics (CAL).

Improvements in the technology associated with the information sciences will, in a broad sense, have their primary potential impact on the distribution of costs, information flow level, information availability, and use among information channels. As technology changes, the information system designer is presented with a continually changing mixture of capabilities and cost levels which form a significant part of his resource allocation decision-making environment. Therefore, he has the responsibility of not only being aware of the state of the art, but of deftly adjusting channel flow and availability based on costs and user needs.

Before discussing potential technological trends and their implications, we need to describe more fully some of the more significant terms used in the above paragraph.

1. Improvements in the technology associated with the information sciences implies both machine and method improvements. Machine includes computers, transmission services, communication terminals, photocomposition or other printing devices, input preparation devices, and facsimile or TV reproduction of images. Method refers to indexing, content representation, and file structure concepts and techniques.

2. Information channels refers to the channels through which or by which information users obtain or receive information. This includes, among others, libraries, symposia, telephone calls, person-to-person discussions, primary journals, secondary journals, reviews, current awareness services, notes, fugitive literature, and manuscripts. The term also implies the possibility of the development of new (now unknown or unused) channels as technological or other changes make them possible.

3. Costs refers primarily to costs to the user but also covers those channel expenses which are partially or wholly deflected by government, foundation or association funding support.

4. Information flow level refers to the quantity of information which can flow through a given channel subject to technological and economic constraints. Emphasis here is on capacity and thus is more significant for some channels than others. For example, the ability to publish journals of several thousand pages each describes high capacity but is not necessarily realistic. However, the ability to transmit the contents of a large data bank from one computer to another would be a better example of the type of capacity referred to here.

5. Information availability refers to how easy it is to obtain information from a channel. This, of course, implies the concepts of perceived accessibility and perceived ease of use.

6. Use refers to the market success of a channel. In various situations this could be measured not only by sales or loan requests but by the amount of time spent by the user in engaging the channel, the distribution and quantity of periods of use, and the perceived and/or reported level of satisfaction.

2. Some Sample Improvements and Their Impact

To illuminate more fully the interactions between technological improvements and channel cost, flow, etc., we can consider several areas.

An improvement implies not only a capability to perform a given function, but the capability to perform that function at a lower cost. Where the lower cost resulting from the improvement reaches a level close to or less than the cost of an alternate method, there will be a tendency to utilize the new method to perform current functions or to adopt new functions heretofore not feasible because of cost or lack of capability. In either case we could have the expansion of a current channel or the creation of a new channel. For example, improvements in photocomposition which make it cheaper to print journals might tend to increase the number of journals. Improvements in computer and communications technology might increase the use of remote terminals for both input and query activities. Improvements in the ability to transfer graphic data via facsimile reproduction could affect both input and query modes and result in new types of services not now feasible.

3. Remote Entry to Multiaccess Computers with Immediate Response

The trend in this area is perhaps the most significant of any of those which will have a strong impact on information systems. Typically it is also burdened with terminological diffusion. This area is usually discussed under the heading time-sharing or on-line systems. To the user all of these terms imply the same type of capability, i.e., access to computer storage and processing services on a fast response basis from a remote location. Before such capability can be economical, it has to be in an environment where many users have access to the same computer.

This technological trend started with the advent of multiprogramming, the residence in computer core memory of more than one program. The Central Processing Unit (CPU) works on (executes) one program at a time, continuing until it is interrupted by, for example, an input/output request. It then proceeds to execute a second program until interrupted, then switches to a third program. A supervisory program which is always resident in one partition of the core memory controls this switching from program to program in addition to performing other required functions.

Multi-programming reduces cost and processing time because it allows programs to be started at microsecond or nanosecond speeds since they are already in core ready to be executed whenever the CPU can get to them. In addition, it allows sharing of other resources for input/output, data handling, etc. The determination of which programs will be executed at a given time is dependent on pre-designated priorities. The program which has the highest priority is normally called the foreground program. Other programs are, as would be expected, called background programs.

An example of foreground/background operation would be a combination of batch and on-line programs. Let us assume we have a language sciences data bank (provided by LINCS) which is accessible by terminals via communication lines. But since the language sciences community of users would probably not be using the computer full time, it would be necessary to use its resources for other jobs. These other jobs could be batch-processing programs which would operate in the background until a request for message processing (an input or a search query, for example) came in from a LINCS terminal. When this request occurred the supervisor program would interrupt the execution of the background programs and give its attention to the foreground program which performed the processing requested by the terminal user. The terminal user would get a response within seconds or less, thus giving him the impression he had the full resources of the computer at his beck and call. Actually, since the computer operates at microsecond speeds and the human user operates at speeds in the range of seconds or minutes the computer is able to handle many requests at apparently the same moment of time. For example, if each request required 10,000 microseconds of computer time it would be possible for the computer to give one-second responses to 100 users who were on line via their terminals at the same time.

The mode of operation described above is normally referred to as time-sharing. Actually, although resources are being shared, time per se is not being shared; instead, there is extremely fast switching from one job or program to another.

Time-sharing is applied to several types of operation. In the example of foreground/background operation given above, a technique called roll-out is used. When a foreground (high priority) program is called for, the background programs can be rolled out of core and the foreground programs rolled-in. This would happen any time the foreground programs required more core space than was available (i.e., not being used by the background programs). Thus, the ability to move programs in and out of core at high speeds is a major feature of a time-sharing environment. Roll-in/roll-out is sometimes called swapping, but the latter term sometimes implies more hardware assistance in the form of hard wired logic networks and registers which enhance the in and out movement of programs.

In some cases several programs have equal priority and are swapped back and forth. When the execution of each program requires a small amount of time which will not delay a response to an on-line terminal by any of the programs, the immediate response illusion can be maintained. However, if one program would require several minutes for execution this would delay responses to terminals requiring the use of smaller faster programs. In order to maintain apparent immediacy the technique of time-slicing is used. This involves setting a maximum limit (in microseconds, milliseconds, or seconds) on the time a given program can be executing. Once this limit is reached it is swapped for another program. Thus, each program operates for a slice of time.

A related technique is paging. Paging is most applicable where large programs which cannot be fitted into the core memory are being executed. If we have a one-half million byte core and a two-million byte program, the program can be divided into four or more pages which are moved in and out of core (usually from a high speed drum memory) as the program is being executed. Paging techniques utilize devices such as associative registers which keep track of status and facilitate the flipping of pages in and out of core. A similar technique called overlapping has long been used by programmers. Computers with paging capabilities in effect do faster and more efficiently (and with less effort on the part of the programmers) what was done in the past via overlay.

Another concept tied in with paging is that of the virtual computer. In the example above, the programmer was dealing with a computer which had virtually two million bytes of core whereas in actuality it had only one-half million bytes. This is somewhat analogous to the situation of the on-line time-sharing user who receives the impression he alone has the full computer resources at his command. The main impact of paging on a user is the ability to design an application as if he were using a computer several times larger than the one he is actually using.

A completely accurate in-depth description of time-sharing is beyond the scope of this paper. The point to be made here is that the developments which have led to this capability have created the opportunity to provide cheaper processing and to provide it for use by remote users.

In addition to cost, a major feature of attraction is the ability to use the power of a large computer without having to make a large investment. The ability to tap large computer power is of major importance to scientific users, but is not as important for information storage and retrieval users. In fact, the basic distinction between scientific and business systems still exists in a time-sharing environment. That is, scientific programs normally involve heavy usage and many iterations of CPU functions, and require relatively minor data input and output. By contract, business programs do not require complex processing, but do require large amounts of data handling and data storage.

A significant question at this point is what improvements are expected in this area? Some relevant technological and growth predictions are.

1. All computing will be on line by 1975.

2. Ninety percent of all computing will be done on line by 1970.

3. Within the next four years, at least some special-purpose time-sharing systems will be capable of handling one thousand simultaneous users.

4. Systems of the future will be largely cathode ray tube (CRT) terminal oriented.

5. By 1970, the majority of computer systems sold will be performing some on-line functions. Huge growth is expected to occur in computers capable of operating on an on-line or time-shared basis.

6. By 1975, computer hardware will become at least one and possibly two or more orders of magnitude cheaper than current systems.

7. The cost per bit of core storage will decrease to less than five per cent of the current cost.

8. However, logic and electronic costs are decreasing at a faster rate than storage costs. (Implications of this trend for manufacturers are the addition of more logic per unit of storage to increase sophistication.)

9. Results of a recent survey indicate a 350% growth in data communications terminals in on-line applications within the next three years.

10. Between 1968 and 1972 there will be a threefold increase in the dollar value of shipments of modems, a fivefold increase in data terminals, and a ninefold increase in concentrators and multiplexors. (Modems, terminals, concentrators, and multiplexors are key equipment items for on-line systems.)

11. Computer costs will decrease drastically relative to common carrier communication costs. (Remote communications are relatively costly and are limited by present facilities.) Some computers will certainly be shared, but the sharing users will generally be within normal toll-free calling distance of the computer (by 1975). Only in those cases where information is volatile relative to average interrogation rates will it be economic to share access to remote storage facilities.

12. Time-sharing technology will evolve into the linking of several or many time-shared (and non-time-shared) computers in networks tied together by common carrier communication channels. The so-called information utility will be a reality by 1975.

13. The threshold of economics for cost justification of common carrier communications line use is a barrier to many users. Communications line costs may decrease with greater use of satellites. On the terminal equipment side when the full impact of the Carterfone decision is implemented, the result should be a lower threshold caused by reduced termination costs. (The Carterfone decision allows users to choose among alternative equipments, instead of being forced to use equipment supplied by the common carrier, thus providing an incentive for manufacturers to develop and attach new, lower-cost devices to on-line networks. Improvements in this area will probably provide us with dramatic cost breakthroughs.)

Although it is difficult to make precise predictions, we can at least be certain that these improvements will bring large scale computer processing within the buying range of more and more users. Thus, it is highly probable that most information services will utilize on-line services. As pointed out above, the on-line capability will normally involve some type of time-sharing. If this is true, our main interest will be in the impact this will have on LINCS.

First we should note the ideal system would handle full text and would perform syntactical and semantic analysis leading to highly sophisticated content representation. Although such capability is not now feasible for operational networks, this does not preclude the economic feasibility of full text search in the immediate future. There are several systems now extant which have unusually good capability for full text search. Probably the best now available is the Mead Data Central system developed by Data Corporation. This system is conceptually quite simple since it involves a variation of the well-known Key-Word-in-Context (KWIC) technique. Each word of text is checked against a "noise" word list containing non-significant words such as articles, prepositions, etc. An index is created from the significant words. Each word is tagged with the document number of the text, the paragraph number, line number within the paragraph, and the sequence number of the word within the line. This inverted file is used for searching. A variety of search techniques could be used. The Data Central system provides typical Boolean search questions, which can be entered via a typewriter or cathode ray tube terminal. Thus, this system can be considered to be both interactive and conversational.

This particular system is mentioned for two reasons. First, it proves that a conceptually simple indexing method provides amazingly good retrieval without the use of statistical association indexing, syntactic analyses, or any of the other schemes for automatic content representation. Second, it may presage the decline of importance of thesauri and of manual indexing, since, with its browsing and interactive capabilities, it to some extent eliminates the stated reasons for these. However, there are some limitations to the system. One is related to the thesaurus rationale, specifically the lack of any word or concept structures or relationships.

Thus, although it is possible to obtain easily and quickly relevant documents or text portions, it does not assure a necessarily high degree of recall. This defect could be corrected to some extent by the use of an internal thesaurus for automatic mapping of terms or, as a less expensive alternative, of a word list with synonyms. The latter list would improve recall and also reduce inverted file size which now can be around 1.3 times the size of the original text file.

Another limitation relates to the scope of systems functions. For example, the Data Central system does not now have the capability to produce high quality secondary journals whether these are in the form of abstract journals or bibliographies with associated subject, author, etc. indexes. (This apparently is planned for the future.) Again, we see here the importance of the distribution of usage of different information channels. If the LINCS project decides to emphasize printed secondary publications this would entail formatting and printing the journals, which would require different file structures. If the on-line interactive channel is selected as the primary LINCS channel, then a system similar to the one described above would receive priority emphasis with a concomitant diminution in the ability to produce high quality indexes. It should be emphasized that it is possible to have both capabilities. Project INTREX includes both but it is not at the operational stage in that its services are not available now as a purchasable service whereas Data Central is available now and is operating now on a real-world basis.

One more statement should be made about the Data Central type of system, viz., the possibility of almost eliminating relevance problems. When searching in an on-line interactive mode, the user continues his search until he has precisely the documents or text portions he wants, which can then be printed on-line via a typewriter or off-line via computer line printer. Thus, when the ultimate user is at the terminal, he can always get 100% relevance (or nothing) with, of course, an unknown degree of recall. Similar results could be achieved with batch searches only after many runs which would not only be problems from the cost angle, but would require days, and possibly weeks, to achieve.

4. Storage Costs

Even though on-line systems are fast becoming economically feasible, most of the capability predictions slide over the cost of storing large amounts of data. One reason for this is that most current on-line systems are either special-purpose (airline reservations, stockmarket quotations, credit card checking, for example), or are primarily for scientific processing where it is usually not necessary to have permanent large data banks. A LINCS data bank consisting of either full text or bibliographic data would, however, require relatively large amounts of storage.

When an on-line system is operating in an interactive mode, disk storage is the normal storage medium used. On an IBM 2314 the cost of this storage is about 15 cents per page. Although newer disk drives are reducing this price, it is still a good figure for comparison purposes. The most obvious comparison is with microfilm, where the cost per page is around one cent. Thus, for the immediate future, it appears microfilm will be a superior storage medium. This fact will tend to make it more feasible for new information systems to adopt microfilm as the storage medium for full text and to use computer storage primarily for indexes of bibliographic items.

However, we still face the question of the number of data banks. If storage is in computer-related media, one central data bank might be feasible. If our full text storage is in microfilm, then we have the problem of remote access to this data bank. The lower costs of this storage medium make it feasible to have many complete or partial collections of microfilm at a variety of user locations. Even when digital storage reaches a competitive cost level, microfilm will continue to be the best answer for older documents.

It should be pointed out that it is possible to have remote access to text in microfilm storage. Project INTREX is now experimenting with this capability. Also, current facsimile reproduction devices can transmit microfilm although normally a hard copy must be made first.

5. Micropublishing

Micropublishing is the offering of banks of information to users on microform to be either viewed or reproduced in hard copy to fit particular user needs. The best known form has been the use of microfiche for full text of documents, usually technical reports, the bibliographic descriptions of which are presented in published indexes for search and reference purposes. The largest micropublishers have been government agencies such as the National Aeronautics and Space Administration (NASA), the Atomic Energy Commission (AEC), the Clearinghouse for Federal Scientific and Technical Information (CFSTI), and the Educational Resources Information Center (ERIC), and University Microfilms of the Xerox Education Division. Companies such as Leasco Data Processing, National Cash Register (NCR), Bell and Howell, and McGraw-Hill have been especially active in this field.

Although dramatic breakthroughs were forecast for microfiche readers and for ultra-microfiche these have not yet reached a cost level making them easily purchasable by individuals. One reason for this is that most development resources have been aimed toward filling market gaps rather than developing competitive products in fields already staked out. However, with the expansion of this total field, the effects of competition should result in reduced relative cost levels for these services in the next three or four years. Lower cost readers would make it possible for more users in the language sciences community to gain access to many relevant documents already available on microfiche from existing information services. Reduced microfiche production costs will make it more feasible for LINCS to provide micropublishing services for language sciences documents not now so available.

6. Vertical Integration of Channels

In this context vertical integration refers to an integrated system which produces output for more than one channel. Most current mechanized systems are horizontally integrated in that they emphasize processing and production of, for example, secondary publications such as bibliographies and indexes. For a given amount of money a broader subject area can be covered in a horizontal system than in a vertical system. The resources of a vertical system would be allocated among primary journals, current awareness services, abstract journals, etc., thus restricting the range or scope of subject coverage.

The issue of interest here is how vertical vs. horizontal considerations are related to problems posed by technological improvements such as full text search and/or on-line interactive systems. The most obvious implication concerns the input cost of full text, i.e., the cost of conversion to a machine-readable medium. In other words, how can the cost of this conversion be justified in terms of use and user benefits? If it is possible to convert the text of one article for the same number of dollars required to convert one or several hundred bibliographic citations, then why is not the former an obvious choice? The answer, of course, lies in the total system planning during which the total user community needs have been assessed. If this assessment reveals the need for primary publications in a subject area, the cost of conversion can be paid via the use of computer-driven photocomposition devices. Once the data are encoded in digital form on magnetic tapes or disks, they can be used for both printing and full text searching, publication of secondary journals, etc. Thus, we have a rule for systems designers to follow. Basically this rule says: if you place a high value on full text search capability, etc., you should integrate this with printing via photocomposition. The major limitation here is that this approach is only feasible when we are dealing with new documents. It does, however, suggest some answers as to the most advantageous approach for active subject areas where there is both a high publication rate and a sizable user community. The possibilities here are not as bleak as they may appear. A similar problem has been faced by many mechanization projects where it was decided to computerize new material only and continue to use conventional methods for retrieving older material. In most cases this approach has led to surprisingly high user satisfaction.

7. The Distributed Computer Concept

A basic objective of LINCS should be to provide the proper information at the proper place at the proper time. However, when all the information is stored at the central computer site, relatively expensive communications services are required to interact with this central data bank. Therefore, if certain sections of the data bank could be stored where they would be used the most often, the total system efficiency would be increased and communications costs could be significantly reduced.

The decrease in logic and electronic costs (especially reflected in large scale integrated circuits, i.e., microelectronics) has made possible the developrent of low cost mini-computers. With a mini-computer used as a terminal, the remote user could perform searches on a small data bank stored in devices peripheral to the mini-computer.

In effect, a page of the central computer memory would be distributed to each remote terminal. If the desired information was not in local storage, the complete local page could be swapped for another page, containing another section of the main data bank, via the communications network.

We cannot assume the distributed computer concept would be applicable for LINCS in the near future since it is very doubtful the use level would be high enough to justify the costs. The concept does suggest, however, possibilities of tying LINCS into a network which also performed other functions which would allow full usage of the capacity made available by this approach. It also suggests the advisability of looking into other network concepts which involve distribution of data bases and programs to remote locations. The following section covers some of these possibilities.

8. Data Bank and Software Transferability

The language sciences are characterized by many "hyphenated" fields. Although linguistics can be considered the core science, as we go toward the periphery, we encounter a variety of fields such as, for example, biolinguistics, psycholinguistics, ethnolinguistics, sociolinguistics, and computational linguistics. Since these "hyphenated" fields are also parts of other subject areas, some information about documents in these areas is already contained in data banks developed in some other subject areas.

As additional data banks are created, this overlap situation will grow. Therefore, LINCS design should include plans on how to tap other data banks. There are, however, very considerable obstacles to the attainment of this objective. Primary among them are:

1. Lack of equipment compatibility and a resultant difficulty in reading a storage medium created by a different type of computer.

2. Lack of a standard character set. (Character set refers to the set of characters, alphabetic, numeric, special, and the binary code configuration used to represent these characters in digital form.)

3. Lack of compatible file formats.

4. Lack of transferable computer programs.

LINCS is somewhat in the position of the tail wagging the dog in attempting to solve these problems. But some policies can be adopted which will put LINCS in the optimal posture for achieving the advantages of interchange and transferability. Among these would be use of flexible file formats or use of a standard file format, which has already been proposed for national use, for interchange of bibliographic data. Another would be the use of a high level computer language with interchange capability. Even if outright data transferability is not possible, it can be facilitated by rigorous specification of format and by programming practices which exclude the embedding of format descriptions in the programs; varying the format from record to record, i.e., all data fields and records should have the same format; and the embedding of format descriptions in the data.

Although programming languages such as COBOL were originally designed to be transferable, this objective has only recently been realized with the development of a standard COBOL which can be executed on many different machines. The problem of transferability has been studied by the United States of America Standards Institute (USASI) and the Committee on Scientific and Technical Information (COSATI), and is now being explored by the MITRE Corporation. Many other institutions, both private and public, have either developed or are now developing solutions to transferability problems. Thus, the future bodes well for LINCS development policies aimed at maximum transferability.