SDC Experiences with Large Data Bases†

CARLOS A. CUADRA

System Development Corporation, Santa Monica, California

Received September 14, 1974

SDC operates a large-data-base system that permits users all over the United States and in several foreign countries to search very large bibliographic files interactively, by means of a terminal and telephone connection. Developing extensive use of such systems requires not only technical considerations—such as proper selection and handling of data base elements—but also a massive educational effort, to help provide the large user community necessary to share the sizable costs of data base acquisition, file development, and storage. The growing acceptance of on-line retrieval services attests to the success of that effort, as well as to their inherent cost-effectiveness.

Some of the characteristics of the large-data-base system that I plan to discuss have already been described by the previous speaker. On-line systems for the retrieval of bibliographic records tend to be similar in concept, but different in mechanics, because of deliberate decisions related to philosophy of design or philosophy of use. One kind of system philosophy has already been illustrated by a previous speaker. The philosophy underlying SDC's system centers on the general principle that a user should be able to ask any question, or issue any command, at any time, without being limited by rigid computer-imposed sequencing of the interactions.

The type of large-data-base system operated by SDC is illustrated in Figure 1. The system user is connected to the central computer by means of a terminal, coupled through a telephone into a special nationwide communication network. The user is then able to interact with the various data bases available through a time-sharing retrieval program. The system also provides for off-line high-speed printing, at the user's request, and for accounting records of system use.

As a matter of accuracy, one of the data bases shown in this figure—the SCISEARCH data base from the Institute for Scientific Information—was not operational until July, 1974. Even without it, the number of bibliographic records on-line each day in this system is approximately 3.5 million. This represents approximately 3.3 billion bytes of disk storage.

SOME ANCIENT SDC HISTORY

It may be useful to review the origins of this system. Our history of on-line systems seems to be written in five-year bytes. In 1960, SDC developed its first interactive retrieval system, known as "Protosynthex." The system used what is now referred to as a full-text approach, the text being the contents of the Golden Book Encyclopedia. In 1965, we put together the first "nationwide" network, in an experiment sponsored by ARPA (the Advanced Research Projects Agency of the Department of Defense). Thirteen government and private organizations, such as the Central Intelligence Agency, the Defense Intelligence Agency, and Battelle Memorial Institute, took part in the experiment, which involved four hours of daily access to what, in 1965, all of us thought of as a large data base—200,000 bibliographic records on foreign technology literature.

Another five years later, in 1970, a very important system came into view—the AIM-TWX system of the National Library of Medicine. This system was the precursor of NLM's current MEDLINE system. The reason that AIM-TWX and MEDLINE are a part of SDC's history is that we developed the systems for NLM and for two-and-a-half years operated them on our computers. Now, NLM and the State University of New York operate the MEDLINE programs on their own computers.

The AIM-TWX data base was not gigantic, by today's standards, but the AIM-TWX service is important because it provided the most important single impetus for today's awareness and acceptance of on-line systems.

GROWTH OF ON-LINE DATA BASE SERVICES

Much of what has been discussed in this conference has been technically oriented. It seems necessary to say a few words about marketing, because the most challenging thing about large data bases today is the fact that they are expensive: expensive to acquire, expensive to process for on-line use, expensive to store, and, in some instances, expensive to search. The key to operating successfully with large data bases in an on-line system is to have enough users to share these costs. But there is a circular aspect to developing a large user clientele. To increase system use, one usually needs to provide access to additional data bases; but doing so is expensive and, in turn, makes it necessary to find still more users. Further, if one is going to serve a great many users, the computer programs underlying the system must be fast enough to handle a large number of user interactions simultaneously, and they must be sufficiently flexible to be changed, in response to the needs of those users.

One of the previous speakers mentioned a limitation in one on-line retrieval system that precluded printing more than 300 citations in an off-line mode. If the system has been designed properly, changing such a limitation ought to be a fairly simple matter, and as it so happens in the system he was describing, changing an off-line print parameter requires only about five minutes of a programmer's time.

But even with an effective and flexible system, it is not easy to bring a large number of users on-line. Potential users have a number of concerns that must be overcome before they become actual users. Figure 2 illustrates some of these concerns. They reflect, in one sense, a continuum of sophistication about on-line systems. The questions at the top of the list were being asked very frequently a year or two ago, when on-line access to large data bases was just beginning to gain widespread attention. At that time, some of the people who viewed on-line demonstrations were convinced that the systems did indeed work, in the hands of

† Presented in the "Conference on Large Data Bases," sponsored by the NAS/NRC Committee on Chemical Information, National Academy of Sciences, May 22-23, 1974.

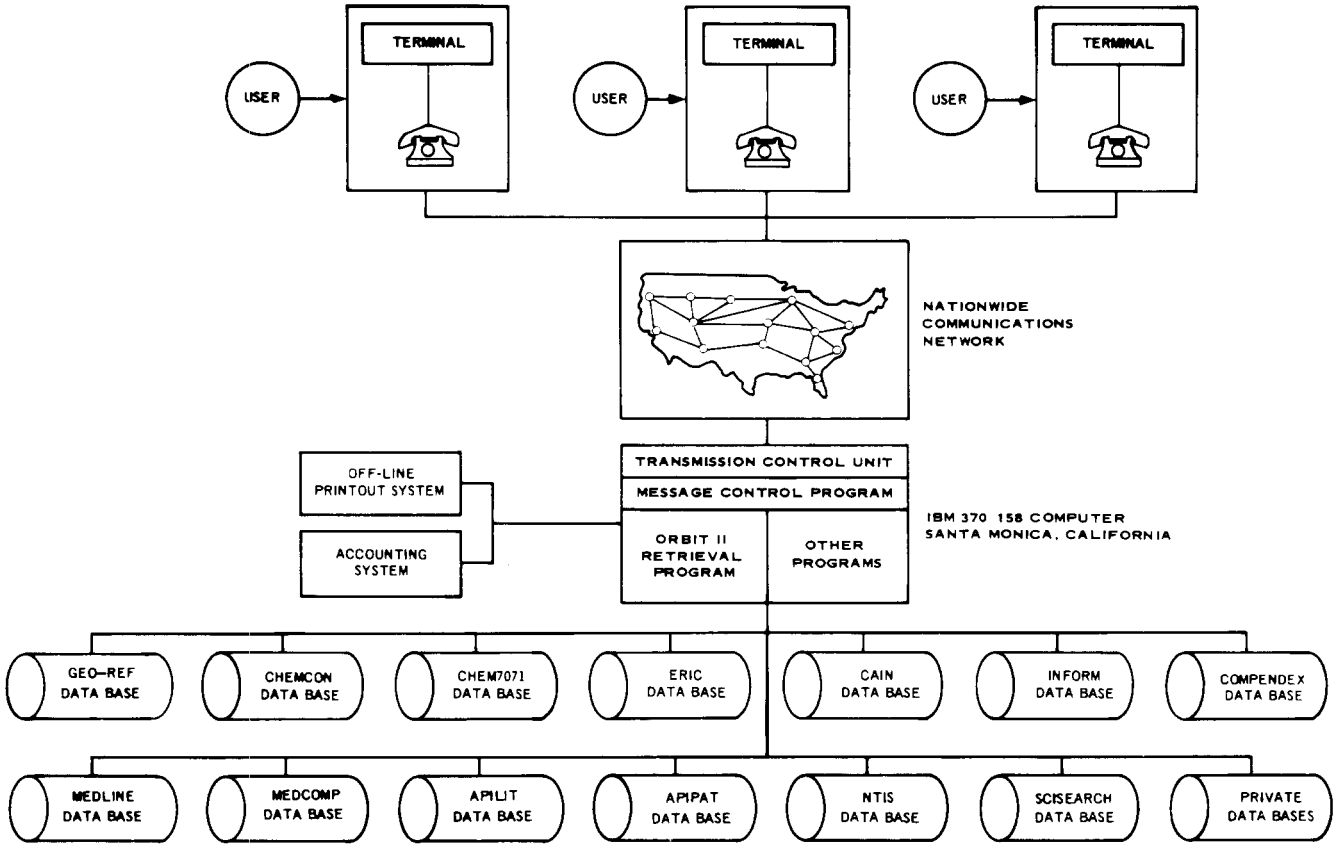

Figure 1. Example of large data base system configuration.

1. Does it really work?
2. Could I make it work?
3. Would it help me?
4. How much would it help me?
5. Might it impact adversely on my present operation?
6. Can I afford it?
7. Can I get the funds for it?
8. What kind of terminal should I get?
9. Which on-line service should I use?

Figure 2. Questions and concerns of the potential user.

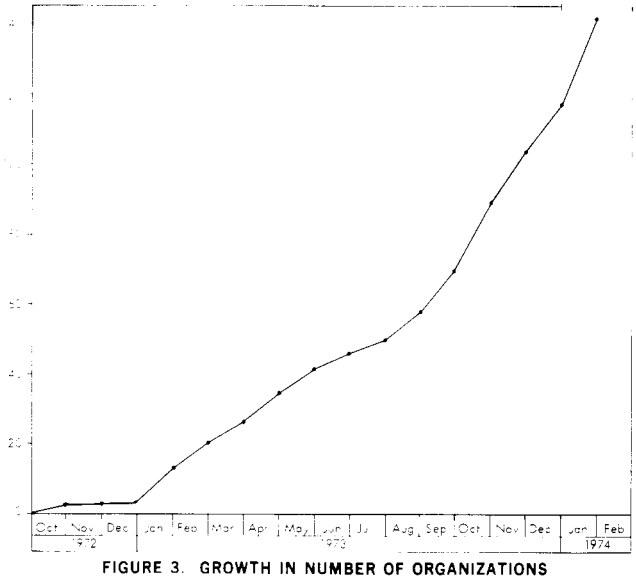

FIGURE 3. GROWTH IN NUMBER OF ORGANIZATIONS

Figure 3. Growth in number of organizations using SDC search service.

systems experts, but they were not at all convinced that they themselves could operate them effectively. Such concerns are tending to disappear, as on-line retrieval becomes a fact of life, evident in many libraries and technical information centers in all areas of the economy.

The last question listed in Figure 2 is not intended to represent the most sophisticated possible question. Libraries and information centers need not and should not think in terms of a single source of information, any more than a homemaker would think of limiting herself to one grocery store or department store. Many organizations may need to use several on-line retrieval systems to meet the varied needs of their users.

There can be no doubt whatever about the acceptance of on-line retrieval services as effective tools. Relatively little data are available on the number of users served by commercial services—these numbers tend to be rather closely guarded—but I can provide, in Figure 3, some data on the growth pattern of SDC's on-line service. Presumably other commercial suppliers can point to comparable growth patterns.

During the first six months after we inaugurated the SDC Search Service, we held our breath. It seemed evident that prospective users of these large data base services had questions at the "will it really work?" end of the scale. Yet the growth in system use was fairly steady. During the next three or four months, there was a change in inflection of the curve: something seemed to be happening. What was happening soon became evident as the growth pattern took on aspects of a veritable boom.

Part of the boom came from word of mouth—from users telling their colleagues that large on-line data base systems not only worked but were cost-effective, too. Conferences focusing on data base systems, such as the present confer-

INITIAL DATA BASES		MEDLINE and ERIC
JANUARY	73	CHEMCON
FEBRUARY	73	CHEM7071
JULY	73	CAIN
SEPTEMBER	73	INFORM
NOVEMBER	73	GEOREF
DECEMBER	73	COMPENDEX
JANUARY	74	APILIT
FEBRUARY	74	APIPAT
MAY	74	NTIS
JUNE	74	SCISEARCH

Figure 4. SDC data base history.

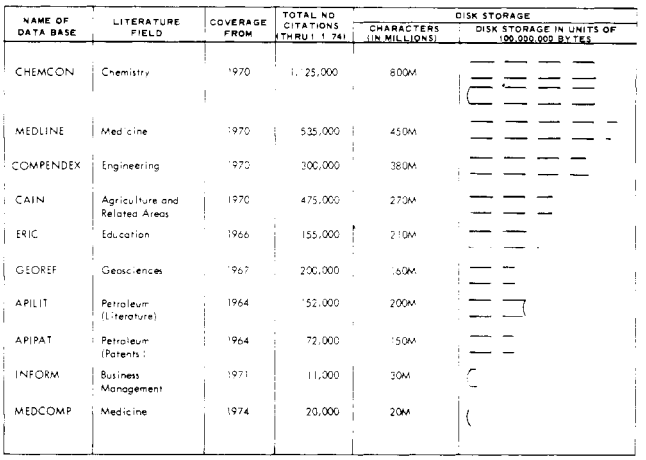

Figure 5. A centralized on-line operation: a measure of the magnitude.

ence, also helped to convince people that on-line systems were, in some sense, respectable and safe.

Another basis for the climbout shown in Figure 3 was the growth in product lines. In our own case, we began with two data bases and 150 million bytes of storage, hardly even respectable nowadays by Dick Giering's standards. From this we have expanded to 12 income-earning files covering 3.5 million records and 3.3 billion bytes of storage. Those are respectable figures by anyone's standards.

Figure 4 shows major additions to the product line over the past year. The first and perhaps the most important addition to the SDC Search Service was CHEMCON, our name for the file developed from Chemical Abstracts Condensates. This is the largest and most heavily used file in our product line.

COMPUTER STORAGE

Each addition to the product line causes both some joy and some pain, because storage is fairly expensive. One must plan his storage requirements carefully, because the faster kinds of random access devices we use have long waiting lines—9 to 12 months, at times.

Figure 5 gives some idea of relative sizes of the data bases with which we are dealing. The numbers were current in January, but they rapidly become obsolete. We add about 30,000 records a month to CHEMCON (Condensates) alone. Altogether, we add 100,000 records a month to our collection of data bases. Nearly all of the files are updated on a monthly basis. The CHEMCON file is updated every two weeks.

Yesterday a speaker mentioned having to do tricky things to keep his system afloat. One of the tricky things we had to do was to learn how to handle a total of 500 million bytes of bibliographic records when we had only 300 million bytes of storage available. Later we needed to handle 2 billion bytes of records when we had only 1 billion bytes of storage available. The problem, simply put, is that at any given point in time, our users want to access more data bases than we have disk storage to handle. Our solution has been what we call our "window" concept: The working day is divided into three-hour blocks. At the end of each three-hour window, some data bases (on disk packs) are demounted and replaced by others. Some data bases are up three hours a day, some six hours a day, and some 12 hours a day. This solution was administrative, not technical. The effect has been to reduce our storage costs by one-third to one-half.

The speaker from IBM yesterday made me quiver when he talked about security and the problem of dropped disks. We have had only one instance of a dropped disk in 18 months, but it was very unnerving. It happened when one of the eight disks on which our CHEMCON data base is loaded froze to the disk drive and had to be removed with a wrench. The person who removed it set the disk down on the floor while he went a few feet away to get another tool. At just that time, one of our regular, careful, conscientious computer operators, seeing a disk pack on the floor and knowing it did not belong there, picked it up by the handle and started to walk away. The handle, unfortunately, was not on tight and the disk dropped and cracked, with $15,000 worth of computer loading time represented in it. Fortunately, our files are backed up and a day later we were back on the air with an intact file.

While I'm talking about equipment, I should mention the computer configuration used to provide access to our various data bases. As Figure 6 shows, we are currently operating on an IBM 370/158, with a VS (Virtual Storage) operating system, and 3330 Model 11 disks.

The figure also shows that we have transitioned through several computers, operating systems, and storage devices. Our president likes to be first with the best, so we tend to get new equipment and operating system releases early in the game. For example, we had the first computer facility outside of IBM that was using the VS operating system on a 158 for production operations.

Being first with the newest is not always a blessing, and all of the members of my staff hold their breath for several days whenever we change any part of the system. They also lose a lot of sleep during system transitions. With the risk of alienating or at least frustrating several hundred users if the system is not functioning properly when 8 AM rolls around, we nearly always do our system checkout in the evening and the very early hours of the morning.

The good news about the new equipment, from the standpoint of a supplier of large on-line data base services, is the new double-density disks. Our old disks held 100 million bytes of storage each; the new ones hold 200 million bytes each, at a lower cost per byte. We expect to add about 200 million bytes a month for the next year to keep up with anticipated file growth and new data bases.

STRUCTURING THE DATA BASE

The computer aspects are, in some respects, the easiest part of operating on-line data base services. The hardest, as I have suggested, is learning how to make ends meet, if one is not running a government-operated or government-subsidized operation.

The next-to-hardest job has to do with the data bases themselves. Operators of on-line data base services must ask themselves questions like these:

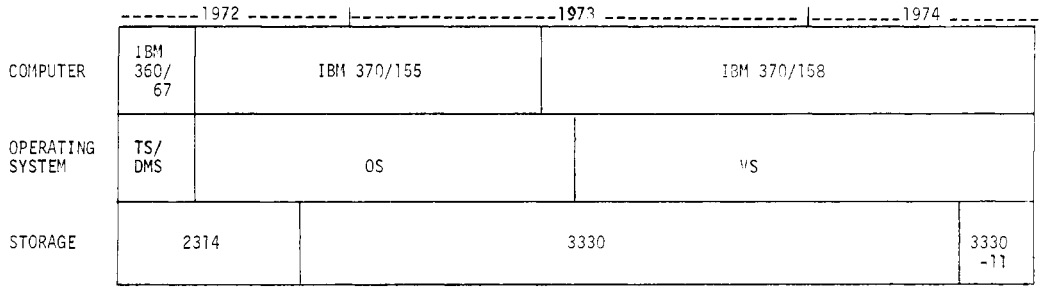

Figure 6. Evolution of computer system supporting SDC search service.

- How large a data base should we acquire?
- How large a data base should we retain or maintain?
- What data elements of each record should we include in the file?
- What access points should we provide for our users?

Deciding on the number of years to place on-line is in part a function of data base size, in part a function of cost, and in part a function of one's estimates of how much one needs to have in order to have a useful and attractive product.

When we decided to put Chemical Abstracts Condensates on-line in December 1972, we decided to start the file from 1970. We also decided to keep 1970 and 1971 in a file separate from 1972-3-4, to make it easier to retire the older records later or in some way reduce their hours of accessibility.

The 1970–71 file receives only about one-fourth of the user traffic that the later file does, but it is still valuable. Interestingly enough, there are some people who will not use our system because it does not go back far enough in time. Even if the Chemical Abstracts Service were willing to provide its early data on tape, beginning with its 1907 files, we could not afford to put them on-line because they could not generate sufficient traffic to cover the costs involved.

Closely related to deciding how much to acquire is deciding how much to retain. Purging is not an inexpensive process on our system, so the question becomes whether to purge or buy additional storage. If the cost of storage were to remain constant or grow, we would probably move toward purging. As it is, the cost for a given amount of storage continues to decline, and we tend to lean toward letting files grow—at least the small- and medium-size files.

One of the most difficult challenges in handling a new data base is deciding which data elements to include in the record. Sometimes the choice is fairly easy. For example, on the INFORM data base, which contains only basic citation information—title, author, source, journal code, and an abstract—there really was no choice. On the other hand, in a file like Chemical Abstracts Condensates, where there are over 30 elements, there were some hard decisions to make. We chose to put 28 of the 30 elements in our CHEMCON data base.

Figure 7 illustrates the fourth of the major problems in structuring the data base: deciding which elements to use as access points. As the figure suggests, the vocabularies contained in the various data bases range from highly controlled to highly uncontrolled. INFORM, for example, has no index terms: any terms that are to be used as access points must be generated by us. We use the title and abstract as the source of these terms. On the other hand, adding uncontrolled terms to MEDLINE, which has a highly controlled vocabulary, tightly linked to a multilevel thesaurus, would add much more noise than good.

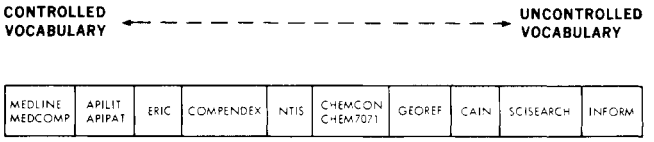

Figure 7. The vocabulary structure continuum.

There is one very important problem associated with large bibliographic data bases that has been particularly difficult to solve: providing full-text copies of the material whose citations have been retrieved from the data base. Although the cost of digital storage keeps getting lower each year, we are not yet at the point where it is economically feasible to maintain the full text of all of the original articles and reports in digital form. At present, it is more reasonable to think of document delivery mechanisms such as the OATS (Original Article Tear Sheet) Service operated by the Institute for Scientific Information.

I am pleased to report to you that SDC and ISI will soon be offering a new type of service that should make the process of acquiring full-text documents faster and more convenient. Beginning in July 1974, users of SDC's Search Service will be able to order full-text copies of documents in ISI's storehouse *on-line*. After identifying the documents of interest, through an on-line search (or by other means), the user can place an order to ISI for the document(s) from the same terminal and using the same system he has been using for searching. Each day ISI will look in its "mailbox," fill the orders the same day, and send them out by airmail.

FINAL COMMENTS

The on-line OATS service provides a good example of new technical developments that are helping to increase the value and importance of on-line data base systems. I am convinced that their use will continue to increase dramatically, and that they will set a new style and standard of living in information service applications.