possible entry points (berkelium, berkelium compounds, inverted entries, actinides, transplutonium, etc.). Only one of the six references was located in the check indicating a reference missed in the original search. Therefore, five references, each of which had berkelium in the title, were not indexed under berkelium in the CA subject index.

The total number of articles common to both CT and NSA was 433 with an overlap between the computer search of CT and the manual search of NSA of 47% (total number of references common to CT and NSA divided by total number of references retrieved by manual search of NSA). The total number of references common to CT, CA, and NSA was 365 or 63% of those retrieved from the three sources (total references in CT, CA, and NSA divided by total references retrieved by computer search of CT).

lated to the transplutonium elements were also cited in CT. However, only about 50% of the CA journal references were retrieved by a computer-based search of CT. In comparison, 47% of the NSA journal references were retrieved by the computer-based search of CT. Over-all, the search questions used in this study retrieved about 40% of the total journal references related to the transplutonium elements found in CA and NSA searches.

## ACKNOWLEDGMENT

Thanks are extended to the staff of the Computer Center of the University of Georgia, Athens, Georgia for their assistance in conducting this study.

## CONCLUSIONS AND SUMMARY

To obtain a complete search of the journal literature of the transplutonium elements for the period 1962 through 1968, it is necessary to search both CA and NSA. This study has shown that the overlap of journal references of the transplutonium elements found in the two reference sources is only about 50%.

A computer-based search of CT, using a profile designed for high precision, recalled approximately 69% of the material cited in the source related to the transplutonium elements. Approximately 59% of this material was retrieved by using the name of the element, while 33% was retrieved using more general terms. The remaining references were retrieved by using another transplutonium element name. About 73% of the journal references in CA re-

## LITERATURE CITED

(1) Ayers, J. B., "The Development of the Literature of the Transplutonium Elements," *J. Tenn. Acad. Sci.* 46, 136 (1971).
(2) Bottle, R. K., "Title Indexes as Alerting Services in the Chemical and Life Sciences," *J. Amer. Soc. Inform. Sci.* 21, 16 (1970).
(3) Chemical Abstracts Service, "CAS Today," Chemical Abstracts Service, Columbus, Ohio, 1967.
(4) Ghiorso, A., Sikkeland, T., Larsh, A. E., and Latimer, R. M., "New Element, Lawrencium, Atomic Number 103," *Phys. Rev. Letters* 6, 473 (1961).
(5) Lynch, J. T., and Smith, G. D. W., "Scientific Information by Computer," *Nature* 230, 153 (1971).
(6) Seaborg, G. F., "The Heavy Elements," *Chem. Eng. News* 24, 1193 (1946).

# Searching the *Nuclear Science Abstracts* Data Base by Use of the Berkeley Mass Storage System*

GLORIA L. SMITH and J. JOANNE HERR**
Lawrence Radiation Laboratory, University of California,
Berkeley, Calif. 94720

**The Berkeley Mass Storage System (MSS) is being used for information retrieval. The MSS has an on-line capacity equivalent to more than 100 IBM-2321 data cells. Advantages of the MSS for information retrieval other than its size are: high serial-read rate, archival data storage, and random-access capability. By use of this device, the search cost in an SDI system based on the *Nuclear Science Abstracts* data base was reduced by 20%. A retrospective search system based on NSA subject categories combines random-access and serial-read search techniques to reduce costs markedly.**

Since 1967, the Information Research Group (IRG) at the Lawrence Radiation Laboratory in Berkeley has been operating an SDI system based on the *Nuclear Science Abstracts* tapes. Although the system is operational and has 120 users, it is not static. The IRG constantly endeavors to

reduce the costs and to improve the quality of the service. Many of the cost-reducing developments have been associated with the availability of random-access devices. The Berkeley Chipstore is one such device that has dramatically influenced the operation of the IRG's system.

The Lawrence Radiation Laboratory (LRL) at Berkeley and Livermore, California has as a part of its computing facilities two of the largest mass storage devices available to date. These are connected to the largest computers made by Control Data Corporation, the CDC-7600 and CDC-

6600. The storage device, the IBM-1360 Photodigital Storage System (called the "Chipstore" at Berkeley), was developed under a special contract by the International Business Machines Corporation. The Livermore system has an on-line capacity of a trillion bits of data; the Berkeley unit is a third as large, and is used chiefly for the storage of particle-tracking data from the Laboratory's high-energy physics program. Detailed discussions of the Mass Storage System (the MSS, consisting of the Chipstore, and associated software and equipment) have been given by Penny et al.,[1] Metcalf,[2] and Kuehler and Kerby.[3]

Recent work by the IRG has shown that the Mass Storage System permits speedy and inexpensive information retrieval from 4.5 years of the *Nuclear Science Abstracts* data base. The properties of the Chipstore, particularly those relevant to its use for information retrieval, are reviewed in this paper, and the IRG's use of the Chipstore for both SDI and retrospective searching is described.

## FILM CHIPS, THE BASIC PHYSICAL UNIT

The smallest physical unit of the system is a $35 \times 70$-mm chip of high-resolution photographic film. Nearly 5 million data bits can be packed onto a single chip. The chips are stored in plastic boxes (about the size of a cigarette pack); 32 chips fit into one box, which can contain up to $1.5 \times 10^8$ bits.

## WRITING, READING, AND STORAGE UNITS

Data from the CDC-6600 computer are recorded on the silver halide film chips by means of a beam of electrons. For data recording, an individual blank chip is positioned in a vacuum chamber, and information is written by repeated sweeps of the electron beam across the chip surface in boustrophedonic fashion (from the Greek *bous*, ox, + *strephein*, to turn—the lines alternate in direction). Binary code is represented by patterns of dark and clear spots. Data are recorded at a rate of more than half a million bits per second; it takes about 18 seconds to write one chip.

The automatic developing unit completes the processing of a chip within 2.5 minutes. Up to eight chips can be at various stages of developing at one time; this overlapping permits the processing throughput to be comparable to the recording unit's throughput. After developing, the chips are checked for errors at a read station. If a chip was not properly recorded, it is discarded, and the data are written onto a new chip.

At a reading station, the requested chip is picked from its box and read with a high-speed, flying-spot scanner at a rate of $1.1 \times 10^6$ bits/second; it takes about 4 seconds to read a full chip. This is more than twice as fast as tapes are read and five times the read rate for an IBM-2321 data cell. The data are read directly from the Chipstore into the user's program in the CDC-6600. The high serial read rate is one of the advantages of the Chipstore for information retrieval.

Another advantage of the Chipstore associated with reading is the archival quality of the storage. Under actual operating conditions, the Chipstore's read-error rate is less than 1/60 that of magnetic tapes. The error rate is low primarily because the data, once written, cannot be changed —either intentionally or accidentally.

The boxes of chips are moved pneumatically between the writer, reader, and box-storage unit (the file). The average time to move a box from the file to the reader is 3 seconds; the time to pick a chip from the box and position it for read-

ing is 0.5 second. The maximum time to fetch a box and pick a chip is 5 seconds. There is some queuing of boxes to be read. Although the box access time may seem long for random-access applications (see below), consideration of the amounts of data accessed indicates that average access times should not be much different from those for data cells.

Boxes of chips can be taken out and inserted into the file manually; this means that the total storage capacity is virtually unlimited. At any one time, there can be up to 2250 boxes on-line; this is equivalent to $3.3 \times 10^{11}$ data bits.

## OTHER COMPONENTS OF THE MASS STORAGE SYSTEM

A small process control computer controls the details of hardware actions—e.g., the physical movement of boxes. A CDC-854 Disk Pack holds all the tables and indexes to the data maintained in the Chipstore. The Chipstore is connected to a CDC-6600 computer, which is multiprogrammed to allow up to 64 jobs to reside in core memory at once and share the central processing unit. One of the 10 peripheral processing units associated with the 6600 controls a high-speed data link between the Chipstore and the 6600.

The MSS software was designed to make interaction with the system easy for the user. A flexible read mechanism allows the user to read either from the Chipstore or from more conventional storage media such as magnetic tapes or disk. A user can label his data with a two-level hierarchy of names for data sets and subsets. These names are stored in the tables on the disk pack, and the user can readily access his data through those tables. The sets and subsets correspond to rather large blocks of data.

The Chipstore's random-access capability is important to its use for information retrieval. Although the system stores tables only for sets and subsets, the user may store more detailed information about his collection—and, with those tables, directly access selected portions of his data.

## STORAGE CAPACITY OF THE CHIPSTORE

A major advantage of the Chipstore for information retrieval is its size.

| Unit | Bits | Capacity Other dimensions |
|---|---|---|
| Chip | $4.7 \times 10^6$ | ⅙ NSA issue |
| Box | $1.5 \times 10^8$ | 5 NSA issues, 1⅓ full reels of tape |
| Chipstore | $3.3 \times 10^{11}$ | 2750 full reels of tape, 110 data cells |

Its size relative to data cells is particularly important. Data cells are often used as the storage devices for on-line retrieval systems, and often the amount of material to be available on-line is limited to one or two data cells' worth simply because data cells are expensive. To give some perspective to this limitation—a data cell would be filled by about 6 years' worth of the *Nuclear Science Abstracts* data base.

The enormous capacity of the Chipstore changes the picture considerably. Those same 6 years of NSA would take less than 1% of the Chipstore's total on-line capacity. Thus, the Chipstore is naturally suited to storing multiple data bases on-line and to storing more information—for instance, abstracts—about each item in those data bases.
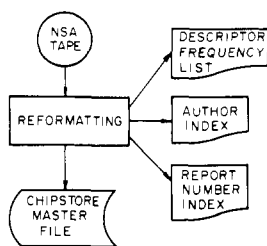
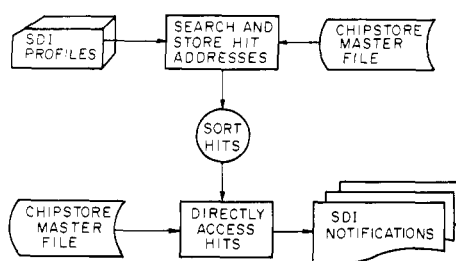Figure 1. NSA tapes are reformatted and copied onto the Chipstore

Figure 2. SDI program does a linear search for up to 200 profiles at a time
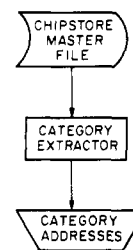
Figure 3. File is inverted on NSA subject categories

## CHIPSTORE ADVANTAGES SUMMARIZED

To review, the advantages of the Chipstore for information retrieval are:
High serial-read rate ($>2 \times$ tape read rate)
Immutability of the stored data (read-error rate $< \frac{1}{60}$ that of tape)
Random-access capability
Large storage capacity ($\approx 110$ data cells)

## IRG USE OF THE CHIPSTORE

Since 1967, the IRG has been operating an SDI system based on the *Nuclear Science Abstracts* tapes. NSA publishes more than 50,000 abstracts each year to give thorough coverage of the nuclear science literature.[4] The NSA tapes, issued twice monthly, give the descriptive cataloging, subject category, and controlled-vocabulary descriptor indexing for each of the items.

In the SDI system, which has been described in detail by Smith et al.,[5] 120 users receive twice-monthly printouts. The profiles average 40 terms each, but the number of terms per profile is essentially unlimited—there is one profile with 1080 terms. The Boolean operators AND, OR, and NOT are used in the searches, and subject categories, corporate authors, journal, country of origin of an item, and language, as well as descriptors, can be used as search elements. Although the system is optimized for SDI searching, the tapes are accumulated for retrospective searching.

## SDI PROCESSING

In the first processing step (Figure 1), the data base is reformatted for more efficient searching and stored on the Chipstore. By-products of this processing are cumulative author and report-number indexes, a printout of the items on the tape for use by the library and the IRG, and a cumulative listing of descriptor frequencies. This last is a very useful tool for profile formulation.

In the processing of the SDI profiles (Figure 2), all the profiles are read into the computer, and the index terms of each item are compared with the terms in each of the profiles. The Chipstore address of each item that satisfies any of the profiles is stored along with the information about the profiles for which it is a hit. After all the data base has been scanned, the addresses of the hits are sorted to make a table of hit addresses for each profile. Then, the hits are directly accessed from the Chipstore master file, and the individualized announcement lists are printed.

In our previous SDI search program, the data base was read from tape, and the hits were written onto disk and the disk addresses stored. After the tape scan was completed, the disk addresses were sorted by profile and the individual printouts were prepared by directly accessing the hits from disk. Because the major cost in this program was in the comparison of terms to determine hits, it was expected that using the Chipstore would not make a large difference in the cost. It was found, however, that the Chipstore's random-access capability reduced the computer cost of an average SDI search by 20%, from 38 to 30 cents per profile.[6]

## RETROSPECTIVE SEARCHING

For retrospective searches more dramatic savings were expected, since the costs of these searches are due in large part to the process of reading the data base. Comparative tests have demonstrated that large cost reductions can be made by use of the Chipstore. A search of 4 years of NSA, carried out in part with tapes and in part with the Chipstore, illustrates the magnitude of the effect. There were 18 terms in the search, and it retrieved about 7 hits per issue. The average cost per issue for the tape searches was $3.53, whereas the average for the Chipstore was $0.90, or about a quarter of the cost of the tape searches.

Retrospective searches can be saved to be run with the multiprofile program—that is, the program used for SDI runs. In this way, the cost of reading the data base is shared among several searches. But we have found that, because most retrospective searches are confined to narrow subject fields, the random-access capability of the Chipstore can be used in a very simple way based on NSA subject categories.

## FILE INVERSION BY NSA CATEGORIES

The NSA indexers assign not only indexing terms, but also subject categories to the items in NSA. We have inverted the file on the 77 NSA subsections (subject categories). This is a rather easy file inversion because there is only one category per item on the tape, the number of categories is small, and the items in any one subsection in a single issue are contiguous (as the subsections are used to arrange the material for the printed NSA). To invert the file (Figure 3), the Chipstore address range for each subsection for each issue is extracted during a scan of the Chipstore master file. This inverted file is similar to a table of contents. To do a retrospective search (Figure 4), the subsections are specified, and the linear search is carried out on only the specified part of the data base.

Table I shows the distribution of the data-base records among the NSA sections; the largest section occupies only 21% of the total data base. A closer examination of this
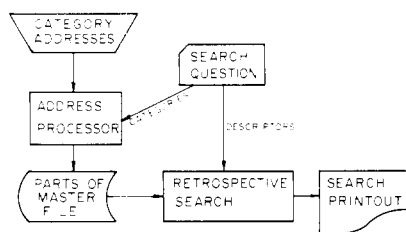
Figure 4. NSA categories are used to select portions of the master file for sequential scanning in retrospective searches

large section, Table II, gives the distribution at the next level of specificity—the subsection—that is available for limiting the retrospective searches. The amount of material scanned can be restricted by section or by subsection, and both are useful for retrospective searches.

A recent search illustrates the cost savings effected by limiting a search by subject category. It was a 41-term search that was limited to the Nuclear Physics section, which occupies 9.4% of the data base. It was carried out on 18 issues and yielded 136 hits. The cost for this search, had it been saved and run with four others on the multiprofile program, would have been $8.30. The category restriction gave two big advantages: it permitted us to run the search immediately (without holding it to combine with others), and it reduced the cost, by 75%, to $1.99.

Table I. Distribution of Data Base among NSA Subject Categories (Sections) Volume 24, Issues 1–24—64,355 Records

| Section Name | Number of Records | Per Cent of Total |
|---|---|---|
| Chemistry | 9,528 | 14.8 |
| Earth Sciences | 1,161 | 1.8 |
| Engineering | 2,717 | 4.2 |
| Instrumentation | 2,812 | 4.4 |
| Life Sciences | 8,405 | 13.1 |
| Metals, Ceramics, and Other Materials | 7,212 | 11.2 |
| General Physics | 13,442 | 20.9 |
| High Energy Physics | 6,299 | 9.8 |
| Nuclear Physics | 6,064 | 9.4 |
| Reactor Technology | 6,563 | 10.2 |

Table II. Distribution of Data Base among NSA Subject Categories; Detail of the General Physics Section

Volume 24, Issues 1–24—64,355 Records for all of NSA; 13,442 Records for General Physics (20.9% of total)

| Subsection Name | Number of Records | Per Cent of Total |
|---|---|---|
| General | 238 | 0.4 |
| Astrophysics | 3158 | 4.9 |
| Atomic and Molecular Physics | 2108 | 3.3 |
| Cosmic Radiation | 264 | 0.4 |
| Direct Energy Conversion | 216 | 0.3 |
| Fluid Physics | 196 | 0.3 |
| Geophysics | 1182 | 1.8 |
| Low-Temperature Physics | 1079 | 1.7 |
| Plasma and Thermonuclear Physics | 2686 | 4.2 |
| Shielding | 273 | 0.4 |
| Solid-State Physics | 1694 | 2.6 |
| Theoretical Physics | 348 | 0.5 |

For a long time, the NSA categories have been used at LRL to restrict both SDI and retrospective searches. Before the Chipstore was used, this kind of restriction did not influence the cost—but it was a very powerful technique for improving the quality of the output by defining the subject area of interest. Now the technique can be used to decrease the cost of the searches strikingly. Of course, the method must be applied with care, but we have found few subject searches that could not be limited in a useful way by category. In this way, the Chipstore's random-access capability can be used to advantage, but without the large cost of inverting the file on the indexing terms. The average cost of inverting one issue by categories alone is $0.78.

## SUMMARY AND FUTURE POSSIBILITIES

Thus far, we have used the faster reads and random-access capability of the Chipstore to reduce the cost of our SDI program by about 20%, and to produce a very-low-cost retrospective-search program simply and without doing a full file inversion. Now there are two routes we could take. We could extend the category gimmick to the SDI service, or we could go immediately to a fully inverted file. Since we now have enough SDI users to recover the cost of a complete file inversion, we are investigating techniques of accomplishing it. With a fully inverted file, retrospective searches will be even less expensive than with the category-restricted linear search, and the restriction will once again be only a device for improving the quality of the output. Moreover, with an inverted file we will be in a position to begin working on an on-line search capability.

## ACKNOWLEDGMENT

## LITERATURE CITED

(1) Penny, S. J., Fink, R., and Alston-Garnjost, M., "Design of a Very Large Storage System," *AFIPS Conference Proceedings* **37,** 45–51 (1970 Fall Joint Computer Conference, Houston, Nov. 17–19, 1970).

(2) Metcalf, M., "The Berkeley Mass Storage System," Lawrence Radiation Laboratory Rept. **UCID-3479,** Sept. 1970.

(3) Kuehler, J. D., and Kerby, H. R., "A Photodigital Mass Storage System," *AFIPS Conference Proceedings* **29,** 735–42 (1966 Fall Joint Computer Conference, San Francisco, Nov. 7–10, 1966).

(4) Shannon, R. L., "*Nuclear Science Abstracts:* A 21-Year Perspective," in "*Handling of Nuclear Information,*" Proceedings of a Symposium, International Atomic Energy Agency, Vienna, Feb. 1970, pp. 379–84.

(5) Smith, G. L., Herr, J. J., and Wakerling, R. K., "An SDI System Based on NSA Magnetic Tapes: User Profiling and the Implications of Decentralized Indexing," *Ibid.,* pp. 251–65.

(6) Herr, J. J., "Comparison of Efficiencies of Various Retrieval Programs on the CDC-6600 Computer," Lawrence Radiation Laboratory Rept. **UCRL-20285** (in preparation).