Chemical Abstracts Service Approach to Management of Large Data Bases†

M. A. HUFFENBERGER and R. L. WIGINGTON*

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received November 29, 1974

When information handling is "the business," as it is at Chemical Abstracts Service (CAS), the total organization must be involved in information management. Since 1967, when, as a result of long-range planning efforts, CAS adopted a "data-base approach" to management of both the processing system and the distribution of information files, CAS has been grappling with the problems of managing large collections of information in computer-based systems. This paper describes what has been done at CAS in the management of large files and what we see as necessary, as a result of our experience, to improve and complete the information management system that is the foundation of our production processes.

In the literature, largely inspired by the concepts in the SHARE/GUIDE documents[6,7,14] on data-base management and attention generated by the CODASYL Data Base Task Group,[5] there has been much discussion of the function and organization of data-base management activities within an enterprise. Most of these publications, however, envision a model for information support to a business where the thrust of the business is something other than information—*e.g.*, automobiles, general retailing, or some other commodity or service. These publications are good sources of techniques and principles, but they tell little about the development and evolution necessary to transform very large information handling operations into a "data-base orientation."

INTEGRATED DATA BASE

The data-base approach asserts that there exists, for each enterprise, an accumulation of information that is pivotal to its operation. This concept implies that the description and treatment of such a collection should not be oriented toward specific processes but should be determined by the value and character of the information itself. An integrated data base usually means an organized collection of computer-readable information in which the information about each entity is recorded once in standardized form, and all access to that information is achieved through indexes and cross-references to the basic record and the authority files that support it.

This definition is sometimes mistakenly thought to imply a single monolithic entity with a completely predetermined set of rules and information structures. However, in real and evolving systems, especially large systems like those at CAS, things are not so simple. It is doubtful that most organizations which handle large data bases would claim to have yet reached such an advanced and uniform state of development or even to desire such a fixed monolithic entity; CAS does not. However, a unified approach and application of data-base principles is leading to effective management of information in the CAS manufacturing system and is beneficially influencing the use of the files that we distribute. Moreover, consistent with the data-base approach, we strongly advocate file-oriented system design rather than process-oriented design which, before the "data base era," guided most of the computer application industry.

CAS FILES

Table I lists the size, activity, and medium for several CAS Master Files. In all, CAS employs approximately 200 named Master Files for various stages in its chemical information and business data processing. This is exclusive of the many transient files (routing files) between processing stages, report printing files, scratch files, test files used in development, and stored information resulting from one-time special processes.

Many CAS files are not only large but they are diverse in their operating characteristics. Some Master Files, *e.g.*, the Chemical Registry System Data Base and the Nomenclature File, are cumulative authority files that are built during processing. Information is then retrieved from them in support of processing and issuance of other files. Some files, like the CODEN/Abbreviated Title File, are relatively static; others, such as the Chemical Registry Files, build at an appreciable rate (currently 300,000 items per year).

Some files, such as the current processing Index Master and Author Index Master Files, fluctuate in size as material continually enters and then is periodically moved to archives when an issuance (*e.g.*, a Volume Index) has been validated. Sizes cited (Table I) for the active portion of these files are simply a snapshot of one point early in 1974 and are not necessarily the maximum size reached just before purging. The annual growth rate cited is an estimate of the gradual increase in average size as the total volume of processed literature expands.

Figure 1 illustrates size variation characteristics for three of the smaller files in the CAS production system. All three types of file growth are shown: sustained growth (cumulative), static, and fluctuating. This graph was drawn from actual data for a three-week period in November 1974. The larger files exhibit the same patterns, although the time constants of variation may be months or years rather than weeks as in the examples shown.

Two large cumulative files for the Eighth Collective Index Period (1967–1971) of *Chemical Abstracts* (CA)—the Index File, including both Chemical Substance and General Subject portions, and the Author Index File—are also listed in Table I. Similar but larger files are now building up for the indexes for the Ninth Collective Index Period (1972–1976).

As shown in the right-hand column of Table I, CAS manages both tape and direct-access files. Through continuing development of the CAS processing system, most

† 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. This work was partially supported by the National Science Foundation under Contract NSF-C656.

* To whom correspondence should be addressed.

Table I. CAS Processing Master Files (Examples of Important Files from 200 Master Files)

	No. of items	Size (May 1974) storage form	Average growth (bytes/year)	Storage medium
Registry Structure Data Base	2.8×10^6	870×10^6	114×10^6	Direct access
Registry Nomenclature File	3.1×10^6	830×10^6	80×10^6	Tape
Index Master File				
Current Processing	2.7×10^6	960×10^6	70×10^6	Tape
Cumulative for 8th Collective Index to *Chemical Abstracts*	$\sim 1 \times 10^7$	$\sim 3.0 \times 10^9$		Tape
General Subject Heading Control File	44×10^3	0.4×10^6	5×10^3	Direct access
Substance Name Match File	3.1×10^6	36×10^6	4×10^6	Direct access
Abstract Publication Data Base	52×10^3	132×10^6	10×10^6	Tape
Headings Master File	130×10^3	115×10^6	8×10^6	Tape
CODEN/Abbreviated Title File	41×10^3	5×10^6	0.1×10^6	Direct access
Author Index Master File				
Current Processing	450×10^3	190×10^6	20×10^6	Tape
Cumulative for 8th Collective Index to *Chemical Abstracts*	3.1×10^6	1.3×10^9		Tape
NCI DR&D CIS Data Base	243×10^3	$<400 \times 10^6$	$\sim 100 \times 10^6$	Direct access

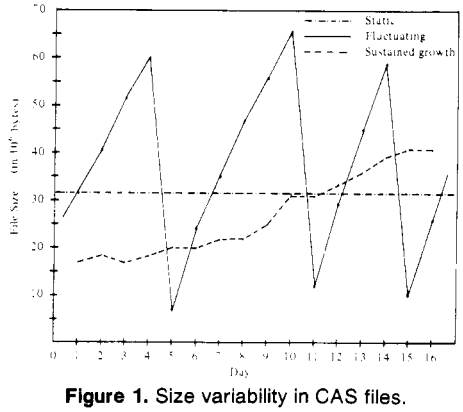

Figure 1. Size variability in CAS files.

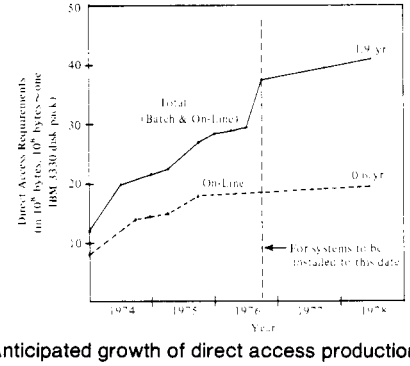

Figure 2. Anticipated growth of direct access production file storage requirements.

active material is being converted to direct-access form, some of it for on-line processing. This conversion from sequential-access tape files to direct-access disk files increases CAS requirements for direct-access storage media. The total direct-access requirements, exclusive of backup and recovery files usually stored on tape, for systems scheduled to be installed between now and mid-1976, are illustrated in Figure 2. Planned installation dates are subject to review as development continues, but the general implications for storage requirements remain, independent of whatever adjustments may be necessary in the detailed schedule.

Figure 2 also shows the on-line production file requirements over the same period. Both curves exclude the storage media needed to support the computer operating system (including virtual storage (VS) paging areas), program libraries, scratch files, test files, and input and output staging areas.

Table II lists the sizes of computer-readable files issued by CAS for use in various search service systems. File sizes reduce somewhat (up to a factor of 2 for tape files) when converted from the internal Standard File Format (SFF)**[2] to the Standard Distribution Format (SDF).[4]

** Standard File Format is a data element storage technique which is designed to provide some measure of data independence. The physical layouts of stored records are inconsequential to the application programs; standard access subroutines and macros are provided for interface to application programs. Data elements may be present or absent and may very in length. They may occur repeatedly. These advantages are possible because SFF records are self-defining; namely, the content, displacement, and length of an occurrence of a data element are stored within the record. The logical records made available to application programs from a single physical record may be very different. These qualities are ideal in a data base environment, and history has shown this mechanism to be an adequate provider. Standard Distribution Format is derived from SFF by simplifying the character set and limiting the storage forms.

This is due to (a) the reduction of the character set for text data from a 2-byte representation to a more compact character set requiring only 1 byte for most characters used and (b) the elimination of some of the data elements used for internal control, reporting, and record keeping. Since these are not useful for searching, they are not forwarded to the issued files. Internal direct-access Master Files use a "compact" Universal Character Set (UCS) (mostly 1 byte/character) for storage form.

SDF is designed for distribution, *i.e.*, transfer between systems, rather than for searching. Most search centers, especially on-line centers, further transform and sometimes further select what to retain in building their own search files.

When all abstracts are handled through a machine system (by mid-1975), the total volume of the issued files—Bibliographic, Abstracts, and Indexes—potentially will be approximately 10^9 bytes/year.

Large data bases imply large systems. Large systems have an inertia that requires the establishment and observance of consistent practices based on general and invariant principles because any system change must take into account the long period of time over which the information has accumulated and will continue to accumulate. For CAS systems, this aspect of design has more implications than any other factor. It places high value on making several careful tradeoff decisions: generality *vs.* specific adaptation, flexibility *vs.* current operating efficiency, and improvement *vs.* consistency with previous practice.

DATA ELEMENTS AND DATA UNITS

The most fundamental aspect of establishing a data base is the definition and design of data elements. Data elements are the bricks from which data bases are built. They

Table II. CAS Issued Files

Product	Period	No. of documents referenced	Size (May 74) SDF form	Growth (bytes/year)
CA Condensates	Mid-1968–1974	1,734,636	1500×10^6	275.6×10^6
CA-Integrated Subject File (CASIF) and CA Subject Index Alert (CASIA)	1967–1974	2,010,552	2685×10^6	483.4×10^6
Chemical-Biological Activities (CBAC)	1965–1974	178,837	444×10^6	83.2×10^6
Polymer Science & Technology (POST)	1967–1974	188,872	413×10^6	85.8×10^6
Chemical Titles (CT)	1962–1974	1,397,350	345×10^6	39.0×10^6
CAS Source Index (CASSI)	1969–1974	34,834	33×10^6	0.2×10^6

are the elementary form in which information is processed, stored, and retrieved. Data elements constitute the standard interface between software modules of various subsystems or between systems and therefore must have a consistent, global meaning, identification number, and format.

Not only are well-defined and uniquely identified data elements important in establishing stable interfaces and to enable reusability of software modules designed to handle them, they also permit powerful, automated information editing. By limiting each edit to a single data element, the known properties and specific conventions can be exploited for validation purposes. Consistency checking for required co-occurrence and inter-element validity is also controlled by the labeled identities of the specific data elements involved.

Some arbitrary and qualitative judgment is involved in establishing data elements, and CAS's understanding of how to do it has grown over the last 8 years. We see increasing overlap between data elements needed in new systems and those used in old systems. This is as desired and is to be expected as automation permeates more and more of our production and administrative operations.

Much as the atoms of Daltonian physics were not indivisible, data elements are not without substructure. Often, it is desirable, for storage and access efficiency and for the ability to construct a specific element from two or more general pieces, to recognize that substructure. To aid both in analyzing the data elements that we have defined over the years and in developing more orderly principles by which to define new data elements, CAS uses the concept of data unit. A data unit is a unique logical and physical datum of a specific type and representation which may be a part of a data element (or a whole data element) and whose specific interpretation depends upon its context when it appears in a data element. The relationship between data elements and data units may be summarized as follows: A body of information can be reduced to a collection of informational quanta called data units which are then recombined into the data processing vehicles called data elements.

Data units can be classified according to a strict interpretation of the information they represent. This technique provides a powerful tool for examining new and existing data elements to determine similarities and uniqueness, based on their constituent data units. Three hierarchical levels are used to classify a data unit: contents code, category code, and unit code.

Most important of the codes is the *contents code*, a strict interpretation of what a piece of data is. It is the primary key to documentation referring to the data unit and to data described by the data unit. An example of a contents code is *Name*. Under this grouping we would find such possibilities as:

Employee Name
Abstractor Name
Company Name
Journal Title

These titles are associated with diverse topics, but the raw data itself remains a name. Thus, in this example, the contents code is independent of specific applications.

The *category code*, a second level of qualification, is a general interpretation of the application of the data and is intended to limit and direct the categorization of a data unit beyond the contents code level. The third level of qualification, the *unit code*, is a specific tag used to complete the classification scheme. Uniqueness of a data unit must be assured by this level of coding. Two examples utilizing all three codes follow:

	Contents code	Category code	Unit code
Example 1	Name	Personal	Abstractor
Example 2	Name	Organization	Company

These examples represent fully classified data units. These data units can be defined as data elements which are composed of a single data unit. The data elements corresponding to the two example data units would probably be titled "Abstractor Name" and "Company Name," with the appropriate descriptions and identification numbers assigned.

A cross-reference table linking classified data units to data elements simplifies the determination of whether a specific data unit appears in any defined data element. For example, newly submitted data elements have their constituent data units classified and checked against the table to prevent redundant data element definitions from occurring.

Data elements, which may be a single unit or a composite of units, can be *functionally* classified into such categories as identifiers (*e.g.*, ID codes such as Registry Number or names such as Index Chemical Name), descriptors (*e.g.*, Chemical Substance Class), or data processing control (*e.g.*, Budget Category Sortkey). Those classed as identifiers account for one-third of all the data elements defined. This is reasonable, since in the CAS system the unique identification of entities—substances, concepts, journals or other publications, authors, etc.—is the basic information handled by an information access system† such as an abstracting and indexing service.

A variety of forms, such as item (a single data unit), array (multiple data unit types), and list (multiple occurrences of one data unit type), are used to combine data units into data elements. The conceptual framework of building SFF file segments from data elements in turn built from data units is illustrated by Figure 3. Figure 4 shows the formalism with which a data element is defined.

Figure 5 shows, in chart form, the growth (between 1967 and 1973) in the number of data elements defined and in the number of data units included in the CAS system. Since data elements may be composites of units, there were, at first (mid-1967 and 1968), more data units than data elements. However, as more data elements were defined, the growth rate in the number of data units slowed. This is shown more clearly in Figure 6.

† An information access system must be able to identify a body of knowledge (document or location in a machine file) that satisfies some need of an information seeker and indicate where the knowledge may be obtained. A total information system requires, in addition, an information delivery function to actually supply the desired information to the requester.

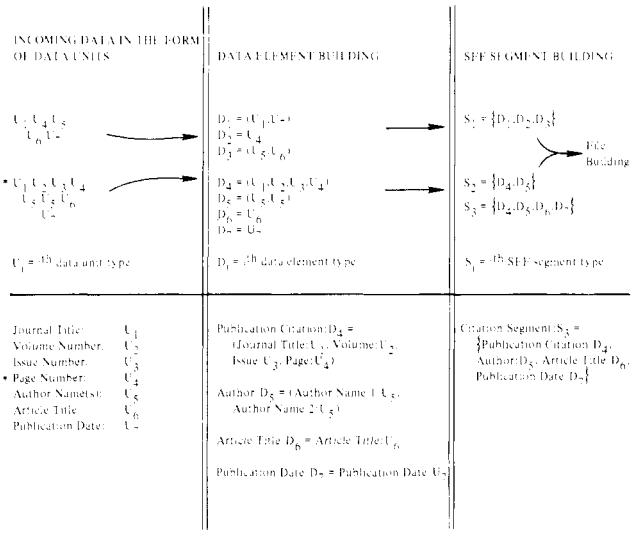

Figure 3. Relationships between data units, data elements, and SFF file segments.

DATA ELEMENT NAME: Publication Citation — ID NUMBER: Ø7 89 ØØ

STORAGE LENGTH (BYTES)	MIN	MAX	TYPICAL	INITIALS
	11	6Ø	31	

DEFINITION

This data element gives the citation of a document within a journal. The citation includes a volume, issue, and page number, along with the journal title.

SYNONYMS

POSITIONS FROM	THRU	STORAGE MODE	COMPONENT DESCRIPTION
1	3	EBCDIC	Volume Number (right justified with leading zeros)
4	6	EBCDIC	Issue Number (right justified with leading zeros)
7	11	EBCDIC	Page Number (right justified with leading zeros)
12	n	EBCDIC	Journal Title (left justified)

Constituent Data Units (by classified title):

Contents	Category	Unit
ID Code	Bibliographic	Volume
ID Code	Bibliographic	Issue
ID Code	Bibliographic	Page
Name	Bibliographic	Journal

Figure 4. Sample data element and its constituent data units.

DATA BASE MANAGEMENT

The successful operation of a data-base-oriented information processing activity cannot be delegated to specialized groups alone. At CAS, all parts of the organization are involved in data-base-related functions.

The content of the chemical information files originates in intellectual analysis performed by staff in the Editorial and Bibliographic Support Divisions of CAS. Accordingly, the semantic definitions of data elements originate in those divisions, and they also establish the content of the authority files used in controlling and editing the information processed. In the same way, appropriate Business Operations, Finance, Marketing, etc., organizations within CAS perform similar functions for the business files.

The design and implementation of data element formats and editing algorithms, file structures, file management

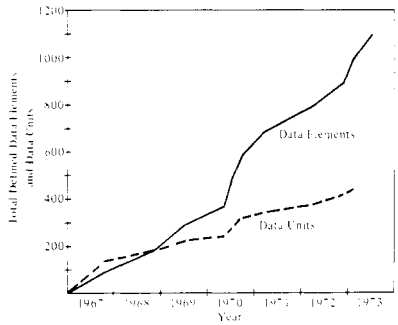

Figure 5. Number of data elements and data units (1967–1973).

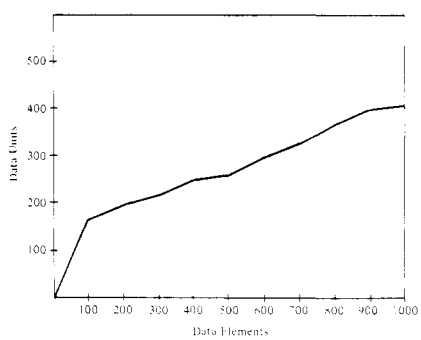

Figure 6. Data units *vs.* data elements.

and access software, and the specific systems for information processing are carried out in the Research and Development Division.

The integration of data base administration, including management of the data sets and archives, is carried out by Data Management Services within the Research and Development Division. This unit's area of responsibility spans new files and systems as well as the continuing production files and systems. Data Management Services also manages the assignment of new data elements, documents all aspects of the data base, performs file reorganizations and recoveries as needed, prepares file control blocks for access software, and maintains the data management software developed by CAS. In addition, it audits the integrity, makeup, and activity of the files and performs analytical studies in support of data-base design and administration.

Data entry and computer processing, including the physical custodianship of storage media, are managed in the Production Operations Division.

ACCOMPLISHMENTS, NEEDS, AND CONTINUING DEVELOPMENT

Using the approaches summarized here, CAS has, over the past 8 years, moved a long way toward establishing a complete data-base management system. Most fundamentally, a standard data element definition and administration approach has been established in all design and production activities. Standard file structures for both tape and direct access files are used. Standard character sets and graphic data structure and the software to handle them are well established.

The software for management of and access to direct access files includes file structure integrity control, access control, and file recovery facilities.[9] A variety of data base auditing and reporting tools have been established and are used in data-base administration.

Many of the internally practiced principles have been carried over into the design and documentation of the computer-readable information distributed by CAS in its Standard Distribution Format.

We see many things, however, that are yet to be done. The integration of processing files into data bases serving multiple functions is only partially accomplished. For example, data-base principles have been applied separately to processing files of bibliographic, abstract, and index information. These files now need to be brought together into a total publication data base; this is the next major step in the evolution of the CAS manufacturing system. Similarly, we plan to integrate the authority files in support of processing. The groundwork for these integrations has been prepared by the accomplishments cited above.

Over several years of development, many specific edit algorithms have been implemented in various parts of the processing system as it has evolved. These also need to be consolidated for efficiency and simplification of processing control.

The computer support of data base and archive record keeping, auditing, and analysis needs to be extended to reduce manual labor and to make data management administration more responsive and totally effective.

CONCLUSION

Based on our experience, the integrated management of large and multifaceted data bases is not something discovered today, adopted as a plan next week, and implemented over the next year. It requires deep understanding before starting, the gaining of working experience with the data bases, and long-term commitment. We have worked at it for approximately 8 years and have made much progress, but full implementation of the data-base principles and practices already known to us will require a few more years. We expect to discover more of the realities as we continue that work.

LITERATURE CITED

(1) Altman, E. B., Astrahan, M. M., Fehder, P. L., and Senko, M. E., "Data Structure and Accessing in Data Base Systems," *IBM Syst. J.*, **12**, 30–93 (1973).
(2) Anzelmo, F. D., "A Data Storage Format for Information System Files," *IEEE Trans. on Comput.*, **C-20**, 39–43 (1971).
(3) Canning, R. G., "The Debate on Data Base Management," *EDP Analyzer*, **10** (March 1972).
(4) "Chemical Abstracts Service Specifications Manual for Computer-Readable Files in Standard Distribution Format," Chemical Abstracts Service, Columbus, Ohio, Aug 1973.
(5) CODASYL, Report of the Codasyl Data Base Task Group, April 1971.
(6) "The Data Base Administrator," Information Management Group, Information Systems Division, GUIDE, Nov 1972.
(7) "Data Base Management System Requirements," Joint Guide-Share Data Base Requirement Groups, 11 Nov 1970.
(8) Engles, R. W., "A Tutorial on Data Base Organization," Report TR00.2004, IBM Corp., System Development Division, Poughkeepsie, N. Y., 1970.
(9) "Facility for Integrated Data Organization (FIDO) User Reference Manual," Chemical Abstracts Service, Columbus, Ohio, April 1974 (available from the National Technical Information Service. Request PB 236020).
(10) Farmer, N. A., Tate, F. A., Watson, C. E., and Wilson, G. A., "Extension and Use of the CAS Chemical Registry System," CAS Report, No. 2, (3–10) April 1973.
(11) Flick, R. A., "Computer Processing of Information for Indexes and Index Files," Fifteenth Chemical Abstracts Service Open Forum, Los Angeles, Calif., March 1971 (available from Chemical Abstracts Service).
(12) "Management of Data Elements in Information Processing," Proceedings of a Symposium Sponsored by ANSI and NBS, McEwen, H. E., Ed., Jan 24–25, 1974.
(13) Nerad, R. A., "Data Administration as the Nerve Center of a Company's Computer Activity," *Data Management*, **11** (10), 26–31 (1973).
(14) "Requirements for the Data Dictionary/Directory Within the GUIDE/SHARE Data Base Management System Concept," GUIDE, 3 Nov 1974.
(15) Rowlett, R. J., and Tate, F. A., "A Computer-Based System for Handling Chemical Nomenclature and Structural Representations," *J. Chem. Doc.*, **12**, 125–8 (1972).
(16) Uhrowczik, P. P., "Data Dictionary/Directories," *IBM Syst. J.*, **12**, 332–349 (1973).