PART I

STANDARDS FOR THE PRESENTATION OF DATA

https://doi.org/10.1017/S0252921100052799 Published online by Cambridge University Press

STANDARDS, MANAGEMENT AND SECURITY OF ASTRONOMICAL DATA SETS

M. S. Davis

Department of Physics and Astronomy
The University of North Carolina at Chapel Hill

INTRODUCTION

Astronomers have historically been among the first to apply or develop, at times, new technologies in the furtherance of their science. This has been especially true in the use of computers from their archaic, antediluvian beginnings to the present highly-developed, time-sharing, multiprocessing, teleprocessing systems. Thus, among the earliest applications in the last half century was the use by L. J. Comrie at the Greenwich Observatory of Hollerith machines for the construction of tables (Comrie, 1928). In 1940 Eckert described punched card methods for numerical integration, computation of a numerical lunar theory, computation of planetary perturbations, as well as applications in photometry and construction of star catalogues (Eckert, 1940). Indeed, one of the earliest collections of files for general use was available at the Watson Astronomical Bureau and included Boss' General Catalogue of 33,342 stars, A. G. Catalogues, Yale Zone Catalogues, and Kohlschütter's Catalogue.

No doubt the Space Age, ushered in by the first Soviet sputnik in 1957 along with the evolution of computers occurring then, heightened the need for extensive data files, at least in some areas. For example, the Yale Catalogues were used as a data bank in computer programs which reduced observations of artificial satellites necessitated by the extremely large number of observations on a growing number of satellites, and required on short time scales for orbit correction.

This Colloquium is testimony to the growth and spread of astronomical data files in recent years to virtually every field

C. Jaschek and G. A. Wilkins (eds.), Compilation, Critical Evaluation, and Distribution of Stellar Data. 3-17.

https://doi.org/10.1017/S0252921100052799 Published online by Cambridge University Press

of astronomy from the earth, to the solar system, to stellar data, to the galaxy, to external galaxies, including reference material like tabulations of atomic spectral data, ephemerides, optical, radio, and x-ray observations. The use of these materials in a variety of fields has more and more pointedly revealed problems arising from data sets, most of which have been designed for their own special purposes, without necessarily considering the possible use of this information for other purposes, or, indeed, as part of a repository in large, astronomical, data banks for general use in a large, possible variety of ways.

It is the recognition of a large constellation of problems involving data sets, focussing attention on Compilation, Critical Evaluations and Distribution of Stellar Data that is the raison d'être for this Colloquium of the I.A.U.

ORGANIZATIONS DEVOTED TO DATA BASE MANAGEMENT

Astronomers should be aware of at least three organizations devoted to the subject of Data Bases and their management:

1. CODASYL (Conference on Data Systems Languages) was organized in 1959. Thanks to a number of Task Groups, in particular, the DBTG (Data Base Task Group), CODASYL has produced a number of consequential reports concerning all aspects of Data Bases and their management in a variety of milieux (CODASYL 1969, 1971), and there exists today a staggering literature on the subject. The implementations of the CODASYL reports have used the high-level, procedural language COBOL and, hence, may be a disadvantage from the astronomer's point of view. Nonetheless, the fundamental concepts developed are sound and based upon vast experience over an extremely wide spectrum of applications.

2. GUIDE-SHARE Data Base Requirements Group (GUIDE-SHARE 1970). This group has approached the subject without defining the syntax of languages to be used and has been primarily concerned with developing concepts of importance. There is not surprisingly a great similarity in the ideas and principles developed by both the CODASYL and the GUIDE-SHARE groups. The group making the GUIDE-SHARE recommendations came from 40 diverse organizations representing years of experience in such areas as banking, life insurance, machine manufacture, government agencies, and universities.

3. CODATA (Committee on Data for Science and Technology, founded in 1968), is an organ of the International

> Council of Scientific Unions of which the I.A.U. is a member and which is represented in it. Hence, it is of especial interest to astronomers. (CODATA, 1968-1976)

While, for the most part, the vast literature dealing with Data Bases and their management is largely irrelevant to the problems of interest to astronomers today, much of the experience and many of the ideas are, indeed, invaluable and will be even more so as astronomical data bases are consolidated into data banks which are to be used in sophisticated ways in the future.

DATA BASE MANAGEMENT SYSTEMS

Let us begin with a general Data Base Management System. Figure 1 is a flow chart and description of such a system. For the system to be viable, the Data Base must be reliable and maintainable in the sense that it can be updated and errors in it can be corrected. Furthermore, it must be capable of search processes to satisfy inquiries and requests made by users.

The Data Base Administrator is a person who defines important parameters concerning the System, among which are:

1. The schema. A complete description of the Data Base.

2. Subschemas. Subsets of the schema made available to various users.

3. Security. In the present context security is meant to be protection of the Data Base against unauthorized access and changes in data. The concept of security has other connotations which will be elaborated on soon.

The Data Base Manager is a program used (1) by the Data Administrator to enforce his policies and (2) by users to access and manipulate data. At this point let us more precisely define several concepts:

1. Privacy. Protection against unauthorized access of data.

2. Integrity. Protection against corruption of data.

3. Security. In general, the term shares meanings between privacy and integrity, but most often is taken to be the same as privacy.

Security is generally enforced by operating techniques on the file level during the running of a program and is most often accomplished

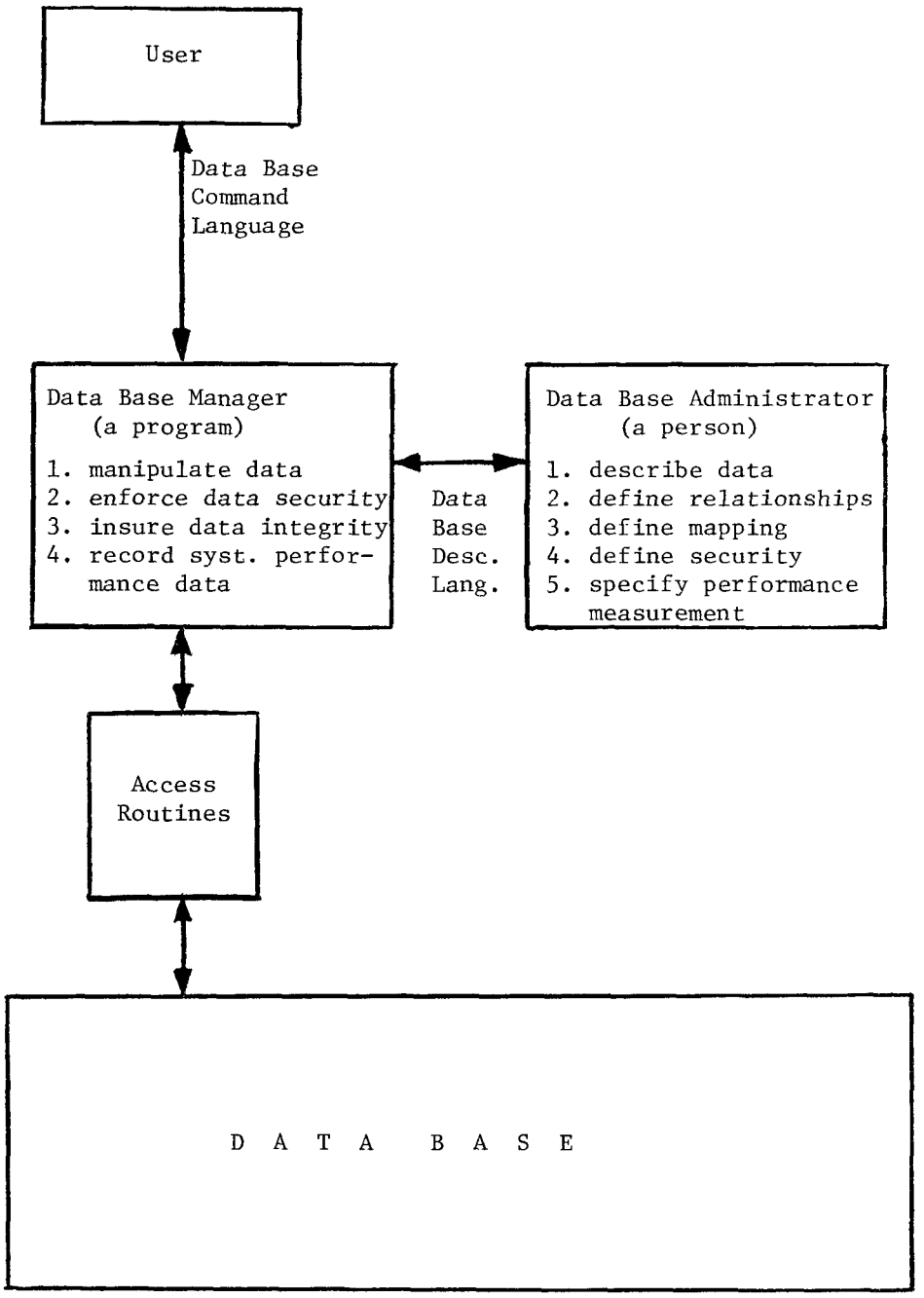
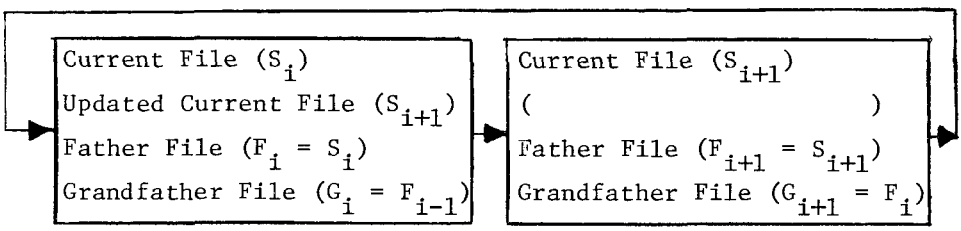


Figure 1. A General Data Base Management System

https://doi.org/10.1017/S0252921100052799 Published online by Cambridge University Press

by making available passwords or privacy keys to privacy locks which are specified in the schema.

Another requirement for viability of the System is Resilience which may be defined as the capacity of a system to recover from errors of systems, program, or hardware types. Of some interest is the re-creation of a Data Base, or parts of it, after corruption or destruction has taken place. If the material in the Data Base is highly volatile (generally, not the case for most astronomical files, but which may apply to some during their creation), one common method for providing backup is periodic dumping of the files, or journalization, which is the logging of data transfers. One of the safest techniques when a file is to be updated, is to generate the new updated file called the "son" file. When this file is deemed correct, the "son" file replaces a previously saved file called the "father" file, the "father" file replaces the "grandfather" file and the "grandfather" file is destroyed. The following diagram outlines the process:



This strategy keeps a copy of the current file (the father file) and the previously updated copy, if it is necessary to reconstruct the current file, or if it is desirable to reconstruct a previous copy of the file. Clearly, older generations can be saved if an extremely high level of safety is needed.

CURRENT ASTRONOMICAL DATA BASES

Most astronomical data presently exist as individual files and reside in a variety of institutions. In 1970 the I.A.U. in collaboration with COSPAR, the Bureau des Longitudes and other scientific institutions established the International Information Bureau on Astronomical Ephemerides (B.I.I.E.A.) whose purpose is to provide information to the international scientific community on availability of astronomical ephemerides, star catalogues, etc. in machine-readable form for use in astronomical and space research. To date 125 information cards have been distributed (BIIEA 1971-1975).

In the context of a large Data Base System with concurrent processing, maintaining integrity is a complex matter involving, as indicated, privacy locks which may be applied to areas,

https://doi.org/10.1017/S0252921100052799 Published online by Cambridge University Press

particularly data sets, records or even items. Furthermore, the extent of such control has implications for efficiency, turn-around and costs.

INTEGRITY OF ASTRONOMICAL DATA FILES

For the most part the astronomical files we are concerned with are independent of large systems, and hence the integrity problem is considerably simpler. Integrity for us is related to the more fundamental questions of the existence and elimination of errors. In a larger sense integrity is also concerned with information which is inclusive enough to make its use meaningful. For example, in a program to reduce observations of artificial satellites, information may be needed concerning the accuracy of the particular catalogue, or a statistical study employing the Yale Parallax Catalogue might need to employ probable errors. These are among the considerations involved in the "critical evaluation of stellar data."

DOCUMENTATION

The insurance of integrity begins with adequate documentation. Documentation for general data description or definition can be prepared on a number of different levels as part of:

1. the sources of the originally compiled data and possibly related procedure manuals.
2. a computer program which processes the data.
3. the computer files.
4. a computer program which manages the file.
5. a printout program.

These can be implemented in a number of ways (London, 1974).

Again, since most astronomical files are standalone and of a historical nature in that they represent parameters measured or calculated at particular times, the most appropriate documentation is that which makes the schema a part of the data files themselves. I would urge general adoption of this idea for all astronomical files. It is an extension of the self-documentation which all well-documented programs have built into themselves. The schema would be written in any natural language approved by the I.A.U. and would not only describe each data item but would provide information on accuracy, errors, caveats, equations used,

in short, the kind of information contained in the introduction to catalogues. This would make each file totally independent in the sense that it completely reproduces the original catalogue. Preferably, I.A.U. conventions should be adhered to in notation and usage.

These goals may not be attainable for many historical files of astronomy but fortunately most of the files of interest to us are already in a quite satisfactory format for uses and applications. The problems I have alluded to earlier will become real to the designer of large astronomical Data Base Management Systems. In fact most users will design their own standalone programs to process the data they require from the files.

ERROR CORRECTION

Thus, the outstanding problem of integrity is the correctness with which the machine-readable media have captured the original source material. A second-order reliability on the correctness of the files, is correction of data in the files discovered to be fallacious in the source material. The first-order of correctness involves the process of "proofreading" in some form or other. Techniques in common use are:

1. proofreading of printouts (for greater safety at least one person to read the manuscript and at least one to read the printout),
2. comparing two or more files prepared independently from the same source material on the same or different media,
3. checking consecutivity of appropriate fields,
4. checking alphabetic or collating sequence order of appropriate fields,
5. checking for blanks, zeroes, etc.,
6. parameter ranges,
7. relational tests,
8. calculation (of calculable quantities) and comparison (as, for example, precession in star catalogues),
9. validity tests,
10. differencing, summation tests.

https://doi.org/10.1017/S0252921100052799 Published online by Cambridge University Press

Some second-order corrections may be found by the above methods. Others are often determined by the original compiler during revisions, updating, and other accesses to the file. Still other corrections are discovered by users during applications when unexpected values or residuals appear. Whenever such errors are discovered it is essential that the user communicate them to the original source and to the repository whence he received the file.

As an example, the Astronomisches Rechen-Institut (ARI) has its machine-readable files in agreement with the printed catalogues unless otherwise stated. Errors which are tabulated in published errata lists have been corrected, as well as those discovered by the ARI and those communicated to it. These errors are published from time to time and the ARI requests users of its files to notify it if additional errors should be found.

The importance of data as error-free as possible, as well as discussions of systematic and random errors can hardly be overemphasized. Macdonald raised the question just a few years ago as to whether most data are worth owning (Macdonald, 1972). While chairman of the Numerical Data Advisory Board of the National Research Council of the U.S.A. he was forced to conclude "No" across the entire spectrum of research to the question posed. His principal reasons were the lack of knowledge about the trustworthiness of the data and often a lack of trust of the measures of uncertainty themselves. Fortunately, astronomers have a long history and tradition of painstaking attention to such matters in most areas of fundamental astronomy and if there is any concern it should be directed at maintenance of high standards of error analysis in the newer disciplines.

STANDARDS FOR ASTRONOMICAL DATA FILES

My remarks so far have dealt primarily with astronomical files of a historical nature comprising most of those currently in existence as ascertained from the list compiled by Wilkins in the Working Group on Numerical Data of I.A.U. Commission 5 and distributed to participants in Colloquium No. 35 or from the cards of the BIIEA. A very few of the files, indeed, are volatile in the sense that data items change, but when they do, it is usually at a very slow rate. Examples of each kind are:

Historical Files	Volatile Files
Astrometric Star Catalogues	Observatories-staff and instruments
Minor Planets	Transition Probabilities

Some of these files are updated with more accurate determinations (such as atomic energy levels, transition probabilities, or

https://doi.org/10.1017/S0252921100052799 Published online by Cambridge University Press

planetary data), or additions (such as double stars, comets, x-ray sources or galaxies).

For this generation of astronomical files then, what should the standards be for compilers of these files? Summarizing what has been said, the standards should be as follows:

1. adherence to I.A.U. notation and conventions,

2. first-order correctness of machine-readable information as the result of scrupulous proofreading of material to ensure precise replication of source material,

3. second-order correctness of the information as the result of a variety of checks on the data itself to discover errors which may exist in the original catalogues or source data,

4. a schema which should be made a part of the file itself to make it totally independent and self-sufficient. The schema should have not only a complete description of data items but should contain all useful information including a discussion of errors, formula used, caveats as to use of the data, discussion of error-correcting techniques used and reliability of the file,

5. if applicable, tests and worked examples should be provided which may be used on the files,

6. periodic, or occasional, publication of errata or corrections of any type, or updating of files, informing users of changes made, or contemplated.

Of particular interest to users of files is the ordering of the records. This, of course, is described in the schema, but special consideration should be given to this feature of a file which may enhance its usefulness greatly for a variety of purposes and is related to making the optimum and most economical applications on the particular computers employed.

Let us define some basic concepts and mention some elementary facts about some of them:

1. Key, sort key, or retrieval key. The identifying field.

2. Collating sequence. The particular order that sequencing follows.

3. Sequential order. Arrangement of a file so that the key field is arranged according to the collating sequence. If a file

https://doi.org/10.1017/S0252921100052799 Published online by Cambridge University Press

organization is sequential, records are stored and accessed consecutively. Advantage - rapid access to next record. Disadvantage - difficulty in correcting or updating the file.

4. Relative random order. According to a particular attribute, the file is in random order, though it is ordered according to the sort key. (For example, the visual magnitudes of stars in the Bright Star Catalog which are arranged by BS=HR numbers are in relative random order.)

5. Random order. The location of a record is random on the access device, though obtainable mainly through (a) dictionary lookup, or (b) calculation (key is mapped onto an address). Advantage - a record may be retrieved in a single access without disturbing other records, and thus, updating or correction of records is easy. Disadvantage - records are normally of equal length. Not rapid for accessing large numbers of records. Overflow problems arise (different records are mapped to same address). Large dictionaries may be necessary.

6. List structure (associative memory). Each record points to (contains the address of) the next record. There are many useful variations of list structures, such as, the ring structure, where the last record points to the first, the coral structure, where there are backwards as well as forward pointers, and the hierarchical structure, where particular attributes have their own pointers through the list or sublists. Advantage - extremely flexible arrangement of memory allowing for changes in list size and updating. Disadvantage - large overhead in managing such a system.

7. Inverted file. The file is ordered according to every attribute of interest which contains pointers to keys (data items become keys). Advantage - extremely useful to extract maximum information from files, particularly when inquiries of the file are unpredictable. Disadvantage - the dictionary may be larger than the data base itself and is difficult to manage and maintain.

Figures 2, 3, and 4 illustrate list and inverted structures using selected information from the Bright Star Catalogue.

PROBLEMS OF COMPATIBILITY

Problems of compatibility exist on many levels, the principal ones being:

1. languages

https://doi.org/10.1017/S0252921100052799 Published online by Cambridge University Press

BS	Name			Double Star Catalogue	RA (2000)	DEC (2000)	Vis. Mag.	Sp. Class	Par
7476	54		SGR	12767A	$19^h40^m44^s$	-16°17'	5.45H	K2III	+".030
7477					19 37 57	+49 17	6.35R	dG6	+.006
7478	12	ϕ	CYG		19 39 23	+30 09	4.64R	G8III-IV	+.007
7479	5	α	SGE	12766	19 40 06	+18 01	4.30R	GOIII	-.004
7480	45		AQL	12775	19 40 43	- 0 37	5.52H	A2	+.014

Figure 2. Exerpts from Catalogue of Bright Stars.

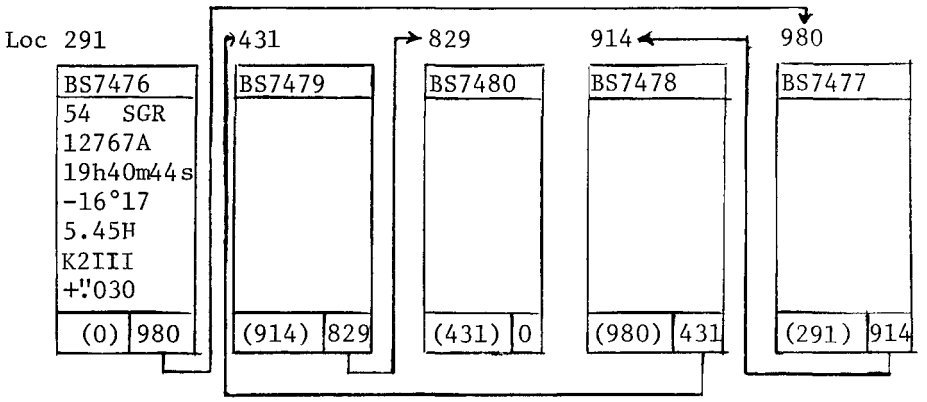


Figure 3. A list structure arrangement of the records in Figure 2. 0 indicates the end of the list. If the 0 were replaced by 291, this would be a ring structure. Backwards pointers are shown in parentheses.

Name	BS
AQL 45	7480
CYG ϕ 12	7478
SGE α 5	7479
SGR 54	7476

Vis.Mag.	BS
$V < 5$	7478,7479
$5 \leq V \leq 6$	7476,7480
$6 < V$	7477

Double Star Cat.	BS
12766	7479
12767A	7476
12775	7480

Spectral Class	BS
A	7480
G	7477,7478,7479
K	7476

R.A. (2000)	BS
19h35m to 19h39m59s	7477,7478
19h40m to 19h44m59s	7476,7479,7480

Parallax	BS
$\pi \leq 0$	7479
$\pi < +.010$	7477,7478
$\pi \geq .010$	7476,7480

Dec. (2000)	BS
> 0	7477,7478,7479
≤ 0	7476,7480

Figure 4. Inverted file structure for the file of Figure 2. Data Items in the Bright Star Catalogue serve as keys in the inverted file.

https://doi.org/10.1017/S0252921100052799 Published online by Cambridge University Press

2. machines or devices.

Language compatibility is, for the most part, a superable problem since the machine-independent development of compiler languages like FORTRAN, ALGOL, COBOL and PL/I. Machine incompatibility, however, has remained a serious problem, often necessitating that data, machine-readable for a particular set of machines, be re-written on media machine-readable on another set of machines. Fortunately, computer manufacturers have moved more and more in the direction of standardization, recognizing the enormous costs involved in re-writing programs and data.

Still, an obvious modus operandi is available in many instances. Clearly, data files stored on data cells, disk packs, drums and similar devices are not generally transferable to other systems. However, most magnetic tapes in use today are 7-track or 9-track, compatible with and fitting the tape drives of most machines. This means, for several reasons, that tapes play one of the dominant roles in data storage and transfer. Data bases on data cells, disk packs, etc. usually have tape media as backup in the event of corruption of the base. Even if they do not, information on the data cells, disk packs, etc. can be written on tape and thus become available to other users.

It used to be that the principal medium for storage was punched cards, and, indeed, many of the astronomical data files stored today in data centers are card decks. This basic medium makes it possible to convert information to media which are otherwise incompatible. In the worst case, where magnetic tape compatibility does not exist, it will still be generally possible to go from tape to punched cards and thence to compatible media. It should be mentioned parenthetically that, in the original compilation of files for a data base, it has been shown to be more economical, as well as producing the least number of errors, to use keytape rather than keypunch devices.

ASTRONOMICAL DATA BASE MANAGEMENT SYSTEMS

In building central astronomical data bases for concurrent use of information through time-sharing or multiprogramming techniques, it will be necessary to have all files resident in the same data base. With the methods described above, it should be possible to have all the files available, even if there are major differences in their structure. It then behooves the management programs to access and manipulate the files so that inquiries and requests can be made across all files. In the language of the CODASYL Report, a "Data Description Language" containing the schema, subschemas, and lock information must be developed for the administrator; a "Data Manipulation Language", which is

procedural, must be developed for access to the Data Base; and, finally, a "Data Management Routine" must be developed to maintain and preserve the integrity of the Data Base.

In operating an Astronomical Data Base Management System, a new class of problems will be encountered, such as "deadlock", when two or more run units are queued and each competes for the unit held by the other, but the considerable experience of others in the field of Data Base Management Systems will make the transition to such a system relatively smooth.

REFERENCES

Bassler, R. A. and Logan, J. J.: 1973, The Technology of Data Base Management Systems, College Readings, Inc., P. O. Box 2323, Arlington, Va., 22202.

Bureau International D'Information sur les Ephémerides Astronomiques, Fiches d'Information N°1 à 125, Palais de L'Institut, 3, rue Mazarine, Paris VI°, France.

CODASYL: 1969, Data Base Task Group Report, October 1969 (out of print).

CODASYL: 1971, Data Task Group Report, April 1971, ACM, 1133 Avenue of the Americas, New York, N. Y., 10036.

CODATA: 1868-1976, CODATA Newsletter 1-16 (twice per year), CODATA Secretariat, 51 Bd. de Montmorency 75016, Paris.

CODATA: 1969, International Compendium of Numerical Data Projects, Springer Verlag, Berlin, N. Y.

CODATA: 1969-1973, CODATA Bulletin 1-11 (irregular), CODATA Secretariat, 51 Bd. de Montmorency, 75016, Paris.

Comrie, L. J.: 1928, On the Construction of Tables by Interpolation, Monthly Notices, R.A.S., Apr. 1928.

Eckert, W. J.: 1940, Punched Card Methods in Scientific Computation, The Thomas J. Watson Astronomical Computing Bureau, Columbia University, N. Y.

Flores, I.: 1970, Data Structure and Management, Prentice-Hall, inc., Englewood Cliffs, N. J.

GUIDE-SHARE Data Base Requirements Group: 1970, Data Base Management Systems Requirements, SHARE Secretary, Suite 750, 25 Broadway, New York, N. Y., 10004.

Hondius, F. W.: 1975, Emerging Data Protection in Europe, North Holland Publ. Co., Amsterdam.

House, W. C.: 1974, Data Base Management, Petrocelli Books, New York.

International Computer State of the Art Report 15: 1973, Data Base Management, Infotech Information Limited, Maidenhead, Berkshire, England.

IAU, Proceedings of the 14th General Assembly: 1971, Proposal for the Establishment of an International Information Bureau on

Astronomical Ephemerides, p. 84, D. Reidel Publ. Co., Dordrecht, Holland.
Katzan, Jr., H.: 1975, Computer Data Management and Data Base Technology, Van Nostrand Reinhold Co., N. Y.
Klimbe, J. W. and Koffeman: 1974, Data Base Management, North Holland Publ. Co., Amsterdam.
London, K. R.: 1974, Documentation Standards - Revised Edition, Petrocelli Books, N. Y., pp. 171-177.

https://doi.org/10.1017/S0252921100052799 Published online by Cambridge University Press