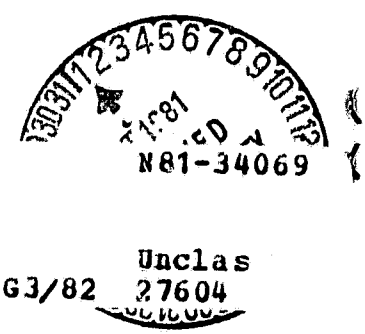
DRAFT



PROGRAM IN INFORMATION POLICY

ENGINEERING-ECONOMIC SYSTEMS DEPARTMENT

STANFORD UNIVERSITY • STANFORD, CALIFORNIA 94305

A SURVEY OF MACHINE READABLE DATA BASES

Peter Matlock

Report No. 34

September 1980
Revised August 1981

National Aeronautics and Space Administration

Contract NASW 3204

PROGRAM IN INFORMATION POLICY

Engineering-Economic Systems Department
Stanford University Stanford, California 94305

1. PREFACE

A major concern of NASA's Technology Transfer Division has been to determine the most effective methods for matching applications of NASA's technology with the needs and interests of non-NASA technologists and scientists. A recurrent theme in the work performed for NASA at Stanford University has been to investigate those methods which allow maximum user involvement in the selection of information deemed relevant to his concerns. This is viewed as important, since it is held that the ultimate "innovator," or "man on the bench" is the best judge of what is of use to him. Thus, minimization of unnecessary pre-screening and pre-selection by third parties better allows the ultimate user to exercise his own judgement, and to make more effective use of the information presented to him.

At the same time, however, it is clear that eliminating or reducing pre-selection of information leads directly to higher user costs, because the amount of information each user must screen is magnified enormously.

Using machine readable data bases with an interactive searching algorithm can be a means for getting information directly to those who will use it, with minimal increase in the user's costs. As an instrument of technology transfer, NASA's RECON files could conceivably play an even greater role in exposing the public to available NASA technology.

With this idea as background, this paper investigates the data base industry, and the functions that NASA already performs as a member of that industry.

2. INTRODUCTION

The growth of machine readable data bases has been rapid over the last fifteen years. Currently over 1000 computerized data bases provide information in all areas of the natural and social sciences, arts and humanities, and business and public policy. A report from International Resource Development estimates revenue from the supply and distribution of on-line data bases at $1.25 billion in 1981, with growth to $5.5 billion in 1991.[1]

This paper describes a sample of the machine readable data bases available to the technologist and researcher in the natural sciences and engineering; and compares them with the data bases and date base services offered by NASA.

The data base industry can be segmented into three categories, following the categorization of Roger Christian.[2] Christian attempted to distinguish three sectors of this industry: the publishers, distributors, and users. The problem with this simple segmentation is that one individual or firm may operate in more than one of these sectors. Distributors publish their own data bases. Publishers market to academic and special libraries, as well as to commercial vendors (who in turn market to academic and

---

[1] Telecommunications Journal, March 23, 1981

[2] Christian, Roger: The Electronic Library: Bibliographic Data Bases 1975-76 Knowledge Industry Publications: White Plains, New York; 1975; page 4

special libraries). Libraries that buy a data base may compete directly with the supplying vendor or publisher. The government sells, buys, and distributes to, from, and in competition with private vendors and publishers. Finally, a publisher's own data base and printed services may compete with each other.

However, despite the often complex flow of goods in this industry, categorizing the data base industry into users, publishers, and distributors is used in this study as a useful classification of the industry.

NASA performs the functions of all three sectors in the data base industry. NASA publishes its own data bases, called the RECON/NASA files. NASA not only compiles much of these files, but publishes them in machine readable form, and distributes them to its user group through the Industrial Applications Centers (IAC's), or the State Technology Applications Centers (STAC's), or by direct dial-in access to NASA's computer center in Maryland. Finally, NASA is itself a subset of the entire user set. Much of the information on the RECON/NASA files is primarily of use within NASA, and the files are widely used by in-house researchers and scientists.

Non-NASA users have access to the RECON files either through the IAC's and STAC's, or through an arrangement whereby prime contractors with NASA may have direct on-line access to the RECON files.

3. FUNCTIONS ASSUMED BY NASA IN THE DATA BASE INDUSTRY

In the machine readable data base industry, NASA not only serves the functions of publishing, distributing, and using its own data bases, but was responsible for developing RECON. RECON is an interactive index and text searching system, which was the precursor to Lockheed's current DIALOG system. Through RECON, users can search several types of data files. These files are listed in Table 1, which gives the number of entries as of July 1, 1980:

Approximately 70% of the RECON entries are comprised by STAR and IAA. Entries in STAR include: NASA, NASA contractor, and NASA grantee reports; reports of other government agencies, univisities, private firms, and domestic and foreign institutions; translated reports; NASA owned patents and patent applications; and dissertations and theses. STAR covers all aspects of aeronautics and space research and development; related basic and applied research; and applications including earth resources, energy development, conservation, oceanography, environmental protection, urban transportation, and topics of national interest.

Entries in IAA include: periodicals (including government sponsored journals); books; meeting papers and conference proceedings of professional and academic societies; and translations of journals and journal articles. The subject matter is aeronautics and space science and technology.

TABLE 1

DATA FILES ON THE RECON DATA BASE

As of July 1, 1980

Document files

Name	Description	Number of Entries	Percentage of Total
STAR	Scientific and Technical Aerospace Reports	453,000	29.4%
IAA	International Aerospace Abstracts	609,000	39.6%
OSTARE	"Old" STAR	148,000	9.6%
LSTAR	Limited STAR	81,000	5.3%
CSTAR	Confidential STAR	144,000	9.4%
Others	Tech Briefs, some ASRDI	104,000	6.7%
Total		1,540,000	100.0%

Special Files

Name	Description	Number of Entries
CPA	Computer Program Abstracts	2,300
R&DCS	NASA Contract Directory	14,600
RTOPS	Research and Technology Operating Plan	7,000
ASRDI(Fire)	Safety File--Fire	4,000
ASRDI(Cryo)	Safety File--Cryogenics	7,000
ASRDI(Mech/Struc)	Safety File--Mechanical and Structural	900
Tech Briefs	Tech Briefs	8,300

NALNET
(NASA Library Network)

Subject	Availability	Number
Books	NASA Holdings	74,000
Books	MARC Tapes	270,000
Periodicals		7,600

Other files on the RECON system include OSTARE, which is

similar to STAR, lists unclassified documents, and is older than STAR. The R&DCS, or NASA Contract Directory, provides the contract number, technical monitor, center, and principal investigator for NASA contracts. The RTOP file lists current and older RTOP's, which are program plans between NASA Headquarters and the NASA Field Centers. The collection of ASRDI files is referred to as the "Safety File," and lists reports collected by the Lewes Research Center addressing safety issues in each of the listed areas. Finally, the Tech Briefs file lists NASA technology available for commercialization, and contains most of the information in the print version of Tech Briefs. The hardcopy Tech Briefs has traditionally served as a major vehicle for transferring information concerning NASA technology to the private sector, and as result there seems to be greater public exposure to the hardcopy Tech Briefs than to the machine readable version.

Some of these data files, in particular the ASRDI, NALNET, RTOP's, and R&DCS files, may be of little interest outside the community of NASA personnel and associated contractors. However, other files, such as STAR, IAA, CPA, and the Tech Briefs may be of considerable utility to engineers and technologists outside this community.

Historically, there has been public (non-NASA) access to these files through either of two avenues. Those members of

the public who are not prime contractors with NASA may submit a search request with a regional IAC or STAC, and the NASA staff will perform the search. This branch of the Technology Transfer Program offers machine and manual searching, historical as well as Selective Dissemination of Information (SDI)--or Current Awareness--searching, and the support staff to interpret and analyze the results of the search.

Prime contractors with NASA may obtain on-line use of the RECON files under a development program to allow greater numbers of users dial-up access to the RECON system. In the past, the number of direct access dial-up ports that RECON's computer could handle has been severely limited. A new front-end processor and larger computer will greatly enhance capacity. As an example of the advances made in expanding its capacity and response, it is reported that RECON response time is down to about two to two and one-half seconds, as opposed to the approximately 26 second average wait in the early stages of the system.[3]

Although much of the file content of NASA's RECON may only be directly pertinent to those closely connected with NASA, there is a substantial amount of information that may be useful outside NASA. Spinoff 1979 makes the point that

---

[3] Figures based on a conversation with Mr. Bill Brown of NASA.

through the IAC'S and STAC's, a non-NASA inquirer has access to over 10 million documents, and of these, 1.5 million are in NASA/RECON. In all, approximately 15,000 scientific and technical journals worldwide are covered, as are publications from other government agencies.[4]

Thus, NASA's data base answers technical and reference questions for non-NASA users and NASA users and contractors, and also serves as an internal research directory for the community of NASA employees and contractors. In terms of technology transfer, NASA uses machine readable data bases to transmit information to private sector technologists through two avenues. One avenue has been mediated through the IAC's and STAC's, and the other is direct and conditional on the prerequisite of being a NASA prime contractor.

One question which can be raised immediately, in light of NASA's plans of expanding RECON capacity and access, is: to what extent have system limitations led NASA to treat the RECON files as a "private resource?" Limited access to RECON through IAC's and STAC's is an efficient distributional procedure when hardware limitations prevent large numbers of dial-up ports. If a "peak load" of independent users cannot be met satisfactorily by a constrained system, then it is obviously more efficient to

---

[4] Spinoff 1979, NASA's Office of Space and Terrestial Applications; page 112. Figures are for 1979.

force all system demands to be spread out over time. A reasonable mechanism for doing this is to channel all, or almost all system demand through agents who will meet each particular demand sequentially. It appears that the IAC's and STAC's have performed this role quite effectively. As the capacity of the RECON system is expanded, it may well be worth emphasizing the other services offered by these NASA Technology Transfer agents, as the importance of their role as intermediaries to RECON can be diminished.

This matter will be dealt with again, following a discussion of a sampling of the machine readable data bases available to private sector technologists in the non-NASA sector.

4. A SAMPLE OF PUBLICALLY AVAILABLE MACHINE READABLE DATA BASES

A sample of data bases was made to compare the scope of coverage and services offered in the machine readable data base industry with that offered through NASA's RECON. The sample was limited to technical and scientific data bases, as these were assumed to be of greatest interest to private sector scientists and technologists. The comparison was made on the basis of subjects covered, type of information recorded in the data base, source of information recorded in the data base, number of years covered, quantity of information and annual additions to the data base, approximate yearly charges for acquisition or lease of the data base, and associated cservices.

The sample comprised 42 data bases, and was drawn from a master list of scientific and technical data bases. The master list was assembled from Information Market Place 1978-1979, which was selected as principal reference document. Information Market Place is a relatively current and complete international compendium of information sources that are publically available. Appendix A is a copy of the Table of Contents of Information Market Place, which is included to illustrate the document's scope of coverage.

In selecting the information sources to be included in their document, the editors of Information Market Place

followed a narrow definition of what constitutes information products and services. Their definition reads as follows:

> Emphasis is on those organizations which, by the application of advanced technologies, create and gather information and add value by performing one or more of the following operations: organizing it, rearranging it, adding other facts, making it more available, or by converting it into a new medium.[5]

Guided by this definition, the editors claim to have undertaken extensive research to locate all agencies and organizations involved in providing information products and services internationally. Upon identifying these organizations, the editors mailed them questionnaires inquiring into their principal and related products and services. Most of the descriptions in the final directory appear to have been self-selected. Apparently, the questionnaire was composed of a series of descriptions, and lists of pre-selected subject areas which were merely to be checked off by the respondent. Certain phrases and combinations of words were repeated quite frequently, particularly in describing the subject areas of the data bases. It is presumed that this is indeed how the information was collected, and this raises the question of how well the descriptions fit the actual content and services of the data bases and data base companies.

---

[5] Information Market Place 1978-1979, page vii

On one hand, this procedure made all respondents look at the same questionnaire with the same questions and categories, and therefore lends a degree of uniformity to the final descriptions of data bases and data base services. This should aid in the comparability of these descriptions in Information Market Place.

On the other hand, with this procedure's reliance on self-reporting, there is the possibility that each respondent interpreted the questionnaire subjectively, although the degree to which one can impose subjective interpretation to words which are generally well defined through industry-wide use appears minimal. It is more likely that variation in the quality of reported information could occur because of varying individual perceptions concerning the expected utility and likely return to contributing to this directory. One can never be sure to what extent respondents presented an exaggerated image of their data base and data base services while perhaps trying to take advantage from some free advertising, or failed to adequately describe their data base and services because they thought this endeavor one not likely to pay off. One can only speculate on the extent of these effects, if they exist at all.

However, with a high voluntary participation rate among American data base publishers, and with certain indications

that data base services were insufficiently described, it appears that data base publishers perceived the directory to be in their own interest, and that these publishers did not exaggerate the extent of their data base coverage and data base services.

If a data base publisher did not return their questionnaire, the editors of the Information Market Place directory performed their own research and presented only the results of this research in the directory, with a note to that effect. Thus, it was possible to measure the participation rate among publishers who had been contacted by Information Market Place. For American companies this rate was 100%, and for all companies internationally the rate was 67%. As only American data base companies were included in the sample discussed in this paper, one can surmise that American companies saw this directory to be in their own interest, and that non-response of the publishers is not a problem for the sample discussed here.

Evidence that the publishers did not exaggerate the extent of their data base and data base services can be observed in the reported services that publishers provide in association with their data bases. For example, it was reported that 54% of the data base publishers offered "machine searching" as a service, whereas only 12% offered "retrospective searching." It seems illogical that there

would in fact be such a high discrepancy, especially since most on-line searches access several years of data. It seems more likely that the services actually performed by data base publishers were somewhat under-reported, or that there was ambiguity in the descriptive labels used in the questionnaire. In either case, the sample discussed in this paper will present only an incomplete picture of any single publisher, although the overall picture across all publishers should be adequate.

4.1 THE SAMPLE

There is a section of Information Market Place that lists machine readable data bases which are publically available worldwide. A "data base" is defined in this directory as a "collection of machine readable records which are periodically updated and which can be processed on computers with the appropriate software."[6] The editors of Information Market Place described approximately 406 data bases in this section, and these included technical and non-technical, domestic and foreign data bases.

---

[6] Information Market Place, page 36

The sample of technical and scientific data bases was collected in two steps. First, all the technical and scientific domestically produced data bases were selected. This was done by a process whereby the author made two passes through the list. In the first pass the author eliminated all data bases whose descriptions did not include areas related to the natural sciences and traditional technical fields. A second pass was made to provide a higher degree of objectivity to this process. This second pass took advantage of a subject index of data bases which is included in Information Market Place. Scientific and technical subject areas were selected, and the data bases listed under each subject heading were cross-checked against the first selection to ensure that there were no omissions. Appendix B contains a list of the subject areas checked.

This procedure resulted in a list of 128 domestically produced data bases that could be of direct interest to a private sector scientist or technologist. Of course, it is always possible that certain data bases, such as the financial data bases, or indices of industrial production, or even the listings of historic homes, could be of direct interest to these scientists and technologists. It is even more likely that certain foreign data bases, such as the European Space Agency's Space Components File, or the Carbon-13 Nuclear Magnetic Resonance CMR, or Hydromechanics and Hydraulic Engineering, would be of direct interest.

This is an important point, since some foreign data bases, such as Aquatic Sciences and Fisheries, are available on-line through services such as Lockheed's DIALOG. However, as specified before, non-technical or foreign data bases were omitted. An expanded study should attempt to identify other types of data bases which could be relevant, and which play a significant role in the data base industry, as perceived from the American viewpoint.

The working sample was selected from this list of 128 alphabetically listed data bases by taking every third listed data base. There was an early oversight in that the Paris-based Aquatic Sciences and Fisheries was originally included in the working sample. This was replaced by the first entry on the master list, which was the API Tech Index.

Summary information on each data base was then collected from Information Market Place, and this information was supplemented by data from certain annual reports, the publication Computer Readable Data Bases: A Directory and Data Sourcebook, and from Excerpts from Directory of On-Line Data Bases. The basic information collected for each data base included its name, publisher, type of coverage (whether each data entry is bibliographic or non-bibliographic, with a description), fields of coverage, years covered, number of entries as of 1978 and number of annual additions, and where

the data base is available if available on-line. Appendix C lists the names of the data bases in the sample, and the publisher of each data base.

Entry 30 in Appendix C is RECON/NASA, which is listed as available through the Knowledge Availability Center, University of Pittsburgh. The RECON/NASA data base is described as bibliographic (listing citations), with multidisciplinary coverage of engineering, science, and project management. It covers the years since 1962, had 720,000 entries in 1978 with 55,000 annual additions and is not listed as being available on-line to the public.

The description of the RECON/NASA files in Information Market Place differs from that given in Spinoff 1980, because the Information Market Place description is only for the Knowledge Availability Center (the Pittsburgh "IAC"). Each IAC essentially has its own procedure for using the RECON system, with the New England Research Applications Center (NERAC) even producing its own tapes. The description of RECON in Information Market Place is in fact a description of information services the Pittsburgh IAC offers, rather than a complete description of the NASA/RECON system.

This discrepancy is of interest, because it appears that the extent of services available through the IAC's and STAC's, as well as the scope of RECON/NASA, are both being

under-reported in a reference document such as Information Market Place. With this insight into the manner in which data bases were described in Information Market Place, the characteristics of the data bases selected in the sample can be discussed. The results will be examined to assess the nature of publically available scientific and technical data bases.

4.2 RESULTS AND DISCUSSION

Information Market Place classified the data bases as either bibliographic or non-bibliographic. A bibliographic data base was intended to supply full bibliographic information of a published document. A non-bibliographic data base supplied specific data, component specifications, descriptions, or non-published reports. Information Market Place did not supply detailed definitions on the difference between these two types, but an article by Doszkocs, Rapp, and Schoolman discussed the types of available data bases in greater detail.[7]

---

[7] Doszkocs, Rapp, Schoolman, "Automated Information Retrieval in Science and Technology," Science, April 4, 1980, pages 25-30

Their discussion of bibliographic data bases, data banks, and the emerging "knowledge" data bases helps clarify the classifications used by Information Market Place. These authors cited bibliographic data bases as those referencing published literature, and which are used most often to locate an article or document. These tend to be computerized versions of existing indexing and abstracting services. Examples include Engineering Index, Science Citation Index, and Government Reports Announcements Index. Non-bibliographic data bases include what these authors refer to as "data banks" and ""knowledge bases." Data banks contain numeric and analytical data obtained from published literature, and often reference the source of information. Examples include the National Library of Medicine's Registry of Toxic Effects of Chemical Substances and Toxicology Data Bank; and the Laboratory Animal Data Bank. The former two contain toxicological, chemical, and pharmacological data for approximatley 36,000 substances (listed by all their names and synonyms, and including their formula), and the latter provides husbandry conditions and physiological and pathological baseline data for laboratory animal groups, and allows interactive statistical analysis. Data from the former two data bases comes from published literature, whereas data for the latter is obtained directly from participating laboratories. Finally the authors discuss the "knowledge data bases," which they compare to

encyclopedias or textbooks. These data bases, such as the Hepatitis Knowledge Base, represent an analysis and synthesis of available knowledge. These examples help illustrate the difference between bibliographic and non-bibliographic data bases.

When the data bases in the sample were analyzed for the relative number of bibliographic and non-bibliographic data bases, the classification of Information Market Place was used. This classification was generally in agreement with that of Doszkocs, Rapp, and Schoolman. 21 of the data bases in the sample were described as bibliographic and 21 were not. Although the numbers need not be exact due to certain definitional problems, there does seem to be an even split between bibliographic and non-bibliographic data bases.

The data bases were then cross-tabulated by bibliographic versus non-bibliographic coverage and multiple versus single field coverage. A data base was defined as having multiple field coverage if its subject area included more than one separate field, even if those fields are related. A data base with single field coverage had, obviously, a subject area in only one distinct field.

As an illustration of this cross-tabulated classification scheme, consider the following examples: Maritime Research Information Service (MRIS) counted as a bibliographic data base with multiple field coverage, since Information Market

Place listed it as "bibliographic," indexing technical reports, journals, and conference papers. Its listing of multiple fields included marine transportation, pollution, business, chemistry, law, metallurgy, and twelve other categories. Cancerproj counted as a bibliographic data base with single field coverage, as it is described as a bibliographic listing of current cancer research projects, and its coverage is limited to researchers, organizations, and funding sources involved in cancer research, and to descriptions of that research.

A non-bibliographic data base with multiple field coverage was the Total Marketing Analysis Research Service, which is a "full text database" listing contract awards by the Department of Defense, and which covers Aerospace and Aeronautical Engineering, Agriculture and Agricultural Engineering, Biology, Business, Economics and Management, and Electronics and Electrical Engineering. Any of the Cordura Publications, Inc. data bases may serve as an example of a non-bibliographic data base with single field coverage. For example, one of these data bases is called Discontinued Thyristor, which provides engineering and purchasing information about discontinued models of this particular electronic component. Its field of coverage is discontinued thyristors exclusively, providing performance specifications, engineering data, type numbers, and manufacturers.

The results of this cross tabulation are shown in Table 2 Note that the data bases are split evenly between those with multiple versus single field coverage, as well as those that are bibliographic versus non-bibliographic.

TABLE 2

DISTRIBUTION OF DATA BASES

Cross Tabulation:
Bibliographic vs. Non-Bibliographic,
By
Multiple vs. Single Field Coverage

Coverage	Bibliographic	Non-Bibliographic
Multiple Field	17	4
Single Field	4	17

Again, although the numbers may not be exact due to definitional problems, there does seem to be a tendency for bibliographic data bases to cover multiple subject fields, and for non-bibliographic data bases to cover a single subject field.

Perhaps this result is not altogether surprising. One would expect a non-bibliographic data base to serve a well defined (i.e., single subject) field, since by nature it is oriented towards this. Somewhat more surprising is the finding that most bibliographic data bases cover multiple subject fields, as there is no a priori reason for this to

be so. One possible explanation is that bibliographic data bases may serve general research or literature reviewing purposes. For these purposes, a bibliographic data base covering a single subject field could be too narrow to be practical.

As an illustration of this problem it is interesting that the Stanford University Engineering Library's principal librarian found that performing an interdisciplinary search on Lockheed's DIALOG system posed a major difficulty. The problem was that too many of the system files (individual data bases) had to be referenced to cover all the relevant material and subject areas desired. In this example, it could be said that the DIALOG system data bases still do not cover enough fields. Enhancements to the DIALOG system currently enable users to search a small set of files simulatanecously, and this is the first step towards the ability to perform simultaneous interdisciplinary searches on multiple data bases.

This question of the content and structure of a data base could be relevant to NASA, particularly if NASA were to consider aggressively tailoring RECON files to users' needs, as more users have the opportunity for direct dial-up access to RECON. The first step in this process would be to identify those potential users and their needs, and determine if NASA's files are of appropriate content and

structure. Of course, as long as only NASA prime contractors have this access, there is a high probability that no file modifications would be necessary. Most NASA prime contractors and NASA personnel are, after all, in closely related areas of work.

A footnote to this discussion is that seven of the seventeen non-bibliographic single field data bases were produced by Cordura Publications, Inc. These data bases were: Discontinued Diode, Discontinued Thyristor, Elastomers, Interface Integrated Circuits, Linear Integrated Circuits, Microwave Tubes, and Optoelectronics. All of these provided specific technical and production information in electronics and materials. These appeared the most specifically targeted data bases in the sample. Judging from Cordura's 1978 and 1979 Annual Reports and the Value Line Survey, these data bases have been quite successful for Cordura, and are the company's most profitable and promising enterprises. Cordura's data bases are an example of targeted computerized information banks, which are valued enough by scientists and technicians that they are commercially successful.

The source of the information listed in each data base is described in Table 3. Table 3 was derived from the publication, Computer Readable Data Bases: A Directory and Data Sourcebook, and data was available for 22 of the data

TABLE 3

SOURCE OF INFORMATION LISTED IN DATA BASE
BY DATA BASE, IN PERCENTAGES

	Journal Articles	Gov't. Reports	Patents	Monograms, Conference Proc. Theses	Reprints, Conference Papers	Manufac-turer's Catalogs	Press Rpts. Broadcasts Releases	Other
1. API Tech. Index	93	1	0	1	5	0	0	0
2. Agricola	90	0	0	10	0	0	0	0
3. Biosis Previews	94.8	.2	0	5	0	0	0	0
4. CA Condensates	72	2	16	10	0	0	0	0
5. CA Subject Index	72	2	16	10	0	0	0	0
6. Cancerproj	0	0	0	0	0	0	0	100
7. Chemical Abstracts Service Source Index	0	0	0	0	0	0	0	100
8. Chemical Industry News	100	0	0	0	0	0	0	0
9. Clinprot	0	0	0	0	0	0	0	100
10. Conference Papers Index	0	0	0	0	0	0	0	100
11. Defense Market Measures System	0	0	0	0	0	0	0	100
12. Drug Product Information File	98	2	0	0	0	0	0	0
13. Energy Line	44	42	0	3	9	0	1	1
14. Food and Agricultural Chemistry	74	1	15	10	0	0	0	0

15. Geological Ref. File	70	3	0	20	6	0	0	1
16. Maritime Research Information Service	67	25	1	2	2	0	0	3
17. NASA/RECON	0	90	1	1	1	0	0	7
18. Pharmaceutical News Index	0	0	0	0	0	0	0	100
19. Polymer Science and Technology	48	1	50	1	0	0	0	0
20. Science Citation Index	100	0	0	0	0	0	0	0
21. TID Keyterm Index	54	1	30	4	1	4	0	6
22. Transdex	0	100	0	0	0	0	0	0

bases. Of the data bases listed, only Bell and Howell's Transdex was comprised of a higher percentage of government reports than was NASA/RECON. 100% of the Transdex entries were listed as government reports, whereas for NASA/RECON this figure was 90%. Energy-line was a distant third in this ranking, with 42% of its entries coming from government reports.

The high percentage of government reports listed in NASA/RECON set RECON apart from the remaining data bases. 82% of these data bases had 3% or less of government reports, 50% had 65% or more of journal articles, and 27% had 100% "other." Thus, NASA is in a distinct minority in terms of its high percentage of government reports. Whether this is perceived by potential users to be an advantage or disadvantage is unanswered, but is obviously important in terms of "marketing" the RECON system.

Table 4 shows the distribution of data bases insofar as the first year of coverage is concerned: No clear trend by five year grouping is apparent, although there does seem to be a growth by decade (10 prior to 1960, 13 in the 1960's, and 17 in the 1970's).

Table 5 illustrates the distribution of data bases by their number of entries in 1978. The largest data base in the sample was the CA (Chemical Abstracts) Subject Index Alert, with 17,896,000 entries. Other large data bases were

TABLE 4

DISTRIBUTION OF DATA BASES

By First Year of Coverage

First Year of Coverage	Number of Data Bases
Prior to 1960	10
1960-1964	7
1965-1969	6
1970-1974	11
1975-1979	6

the Science Citation Index with 5 million entries, and CA Condensates with 3,133,600.

TABLE 5

DISTRIBUTION OF DATA BASES

By Number of Entries, 1978

Number of Entries	Number of Data Bases
Less than 100,000	22
100,000-500,000	10
500,000-1,000,000	4
1,000,000+	6

The distribution of the 22 smaller data bases (those with 100,000 entries or less) is exhibited in Table 6 Eighteen of these twenty-two smaller data bases had less than 50,000 entries. Figure One plots the number of data bases on a log scale of size. The distribution is

Figure One

Distribution of Data Bases
By Number of Entries, 1978
Log scale for number of entries

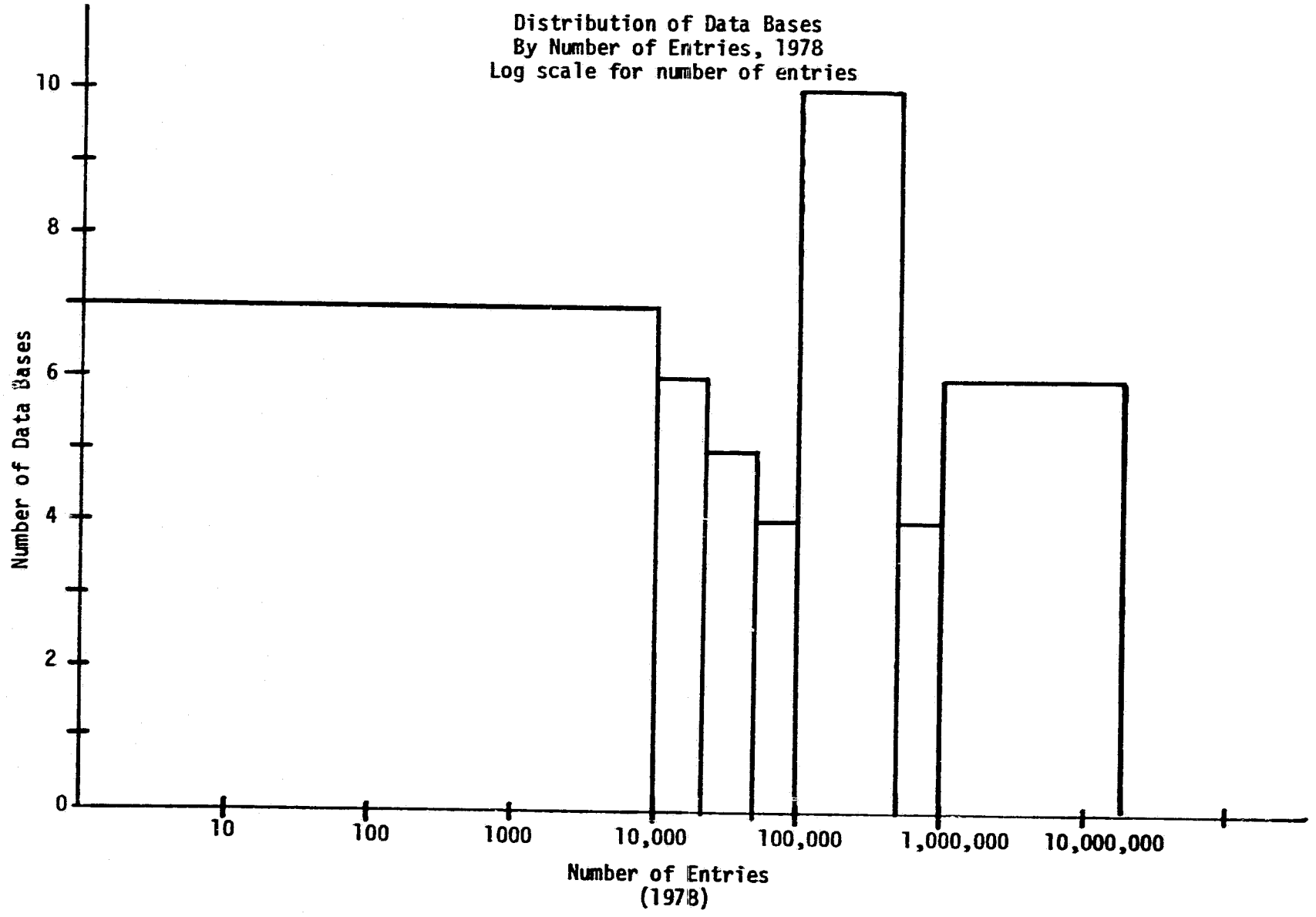


TABLE 6

DISTRIBUTION OF DATA BASES

By Number of Entries, 1978
Less Than 100,000 Entries

Number of Entries	Number of Data Bases
Less than 10,000	7
10,000-20,000	6
20,000-50,000	5
50,000-100,000	4

essentially constant, at a level of approximately six data bases for each size grouping. It should be noted, however, that most of the data bases with less than 10,000 entries were the highly-specialized, non-bibliographic Cordura Publications Inc. data bases, as well as the extremely small ARPANET Requests for Comments (which is essentially an on-line suggestion box).

Table 7 provides the distribution of data bases with respect to the number of annual additional entries:

TABLE 7

DISTRIBUTION OF DATA BASES

By Number of Annual Additional Entries

Number of Annual Additional Entries	Number of Data Bases
Less than 50,000	24
50,000-100,000	6
100,000+	4

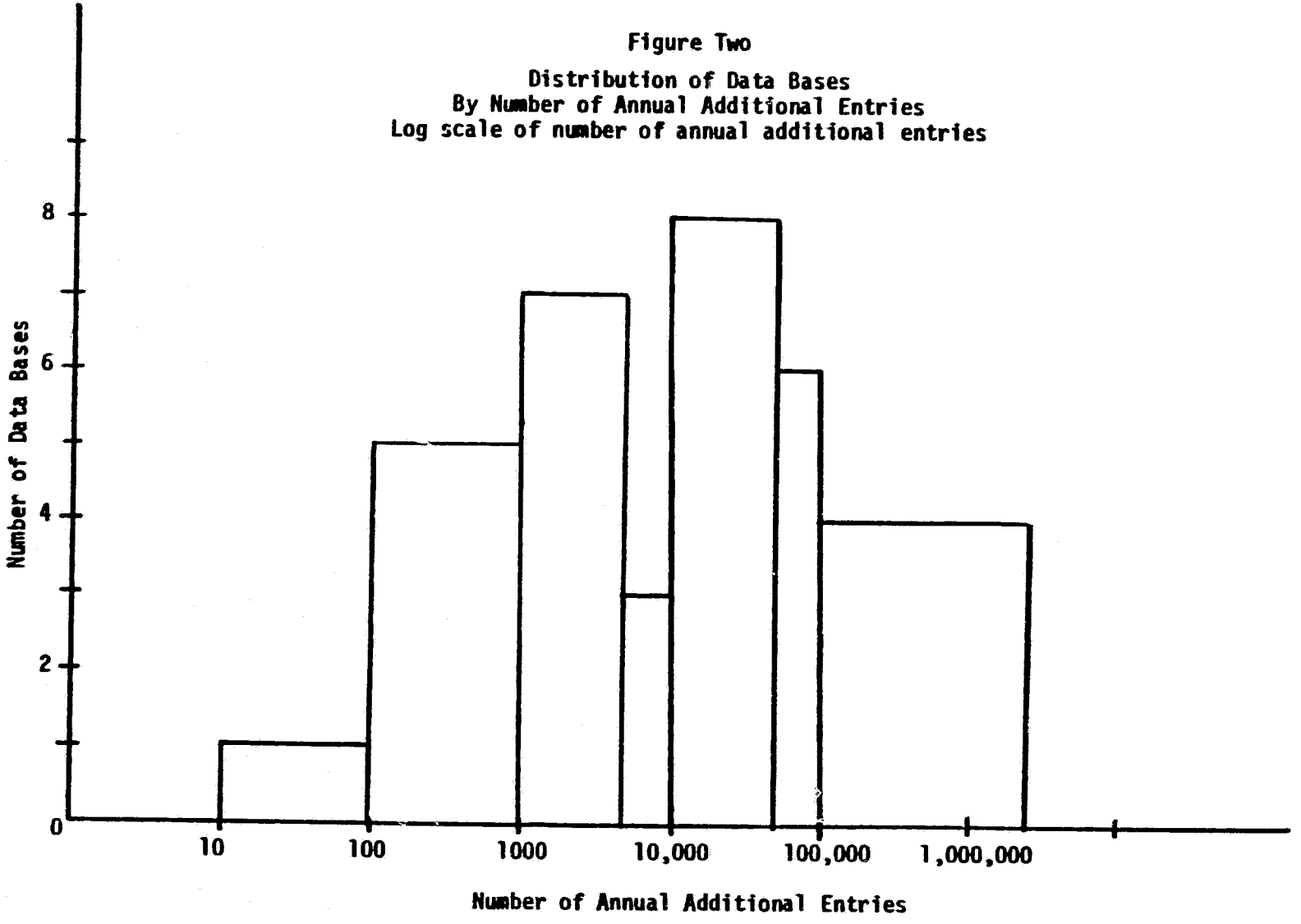
This distribution is represented in Figure Two, which plots the number of data bases on a log scale of the number of annual additional entries. In contrast to the plot of number of entries (Figure One), the plot of number of annual additional entries does not appear constant. The data base with the largest number of annual additions was the CA Subject Index Alert, with 2,600,000 additional entries annually. CA Subject Index Alert was also the largest data base in size as of 1978. Data was not available for every data base.

Data representing the consumer's price to acquire, lease, or license each data base was collected from Computer Readable Data Bases: A Directory and Data Sourcebook. This data was used to try to estimate a first order relation between the "price" charged for each data base and a data base characteristic that in theory would contribute to the cost of producing the data base. In this paper the size of the data base and the number of annual additions to the data base were selected as characteristics hypothesized to most likely affect the price charged to data base users.

The desired result is to test whether price charged for a data base equals the marginal cost of producing that data base. The marginal cost of producing the data base is approximated by a simple cost function with a single data base characteristic as its argument. One would prefer time

Figure Two

Distribution of Data Bases
By Number of Annual Additional Entries
Log scale of number of annual additional entries



series data for each data base to perform this measure, but only a cross sectional sample for one year is available. Furthermore, the "price" variable is not an equivalent measure for each data base. The figures reported were for either acquisition, lease, or license of the data base. For simplicity the acquisition price was treated equivalently to one year's lease or license price, even though an acquired data base would provide additional years of service at no additional yearly charge, but additional years of service from a leased or licensed data base would incur such charges. Finally, only the user's "fixed" costs of gaining initial access to a data base's information were considered. Most of the leased or licensed data bases charged an additional hourly search charge, or "variable" cost to the user. These additional costs were not considered.

The first set of regressions were to determine if there were a relationship between the size of the data base and the price charged the user. Simple regressions were run for the seventeen data bases for which information was available, and then for data bases classified either as private, or non-profit and government. F Tests on each regression showed that none of them were significant at the 90% confidence level. Results are below:

$$y = mx + b$$

y = acquisition, lease, or license costs

x = number of data base entries

All data bases:

(1) $y = .0023x + 3630$

N = 17

$F(1,15) = 3.466$ insignificant

Private data bases:

(2) $y = .0019x + 6031$

N = 11

$F(1,9) = 1.316$ insignificant

Government and non-profit data bases:

(3) $y = -.0007x + 1450$

N = 6

$F(1,4) = .676$ insignificant

Regressions were run with the number of annual additions to the data base as the independent variable, to see if this variable might not be a proxy for marginal costs, and produce a correlation with the data base price charged. There were insufficient data points to do this for data bases produced by government and non-profit organizations,

so the regressions were only performed on all 13 data bases for which there was data, and for the 11 private ones. All regressions were insignificant at the 90% confidence level.

$$y = mx + b$$

y = acquisition, lease or license costs

x = number of annual additions to data bases

All data bases:

$$(4) \qquad y = .0186x + 4322$$

$$N = 13$$

$$F(1,11) = 1.482 \quad \text{insignificant}$$

Private Data Bases:

$$(5) \qquad y = .0171x + 5033$$

$$N = 11$$

$$F(1,9) = .9962 \quad \text{insignificant}$$

The insignificance of these regressions may be attributed to any of the following reasons:

1. Misspecification of variables
2. Errors in the data
3. Misspecification of the model
4. Misspecification of the functional form
5. Lack of any true relationship

Possible misspecification of the variables and errors in the data have already been discussed, and a correction for possible heteroskedasticity (problems with correlating variables from data bases of different sizes) yielded no better results. The simple regressions used here may not reflect all the costing and pricing complexities that actually occur, indicating an inaccurate model. The functional form of these relationships may not be linear, but an experiment with log forms did not improve results. Finally, it is possible that there is no true relationship between these variables, although the argument for this on the basis of current evidence is weak. If there were to be no true relationship, however, economic questions would arise concerning the extent of potential inefficiencies in the still young machine readable data base industry.

Twenty of the forty-two data bases were listed as being publically available on-line, through services such as Lockheed's DIALOG. The most commonly cited on-line service was Lockheed's, but the following were also listed: System Development Corporation's (SDC's) ORBIT, the National Library of Medicine's MEDLINE (MEDLARS on-line), the Bibliographic Retrieval Service (BRS), Triangle Universities Computation Center, General Electric Mark III, and Drilling Activity Analysis System. It would be fallacious to assume

that the twenty-two data bases for which no on-line vendor was listed are not available in some on-line capacity. For example, Cordura Publications, Inc. offers on-line service to its data bases through a contractual agreement. Although this type of service is quite different from the "data base supermarket" service offered by Lockheed,[8] SDC, or DRS, it is still an on-line service.

It may be that some data bases can still be accessed only by batch methods. Further investigation would be required to determine the exact percentage of data bases available on-line. However, it is instructive that in the Doszkocs, Rapp, and Schoolman article cited earlier, the authors note that the majority of the 528 bibliographic data bases they are familiar with can be searched on-line.[9] The authors make no mention of the proportion of non-bibliographic data bases available on-line, but one would expect this proportion to be high due to the nature of these data bases and their use.

---

[8] See Appendix D

[9] Doszkocs, Rapp, and Schoolman: ibid. page 25

5. DATA BASE PUBLISHERS

Information on the publishers of the data bases in the sample was obtained by referencing the "Database Publishers" section of Information Market Place. Twenty-six publishers were responsible for the forty-two sample data bases, and their names are listed in alphabetical order in Appendix E. Information was collected for each publisher, and this information included the total number of machine readable data bases and associated print products produced, and the availability of special training sessions, additional publications, and customized services.

The twenty-first publisher listed in Appendix E is the NASA-IAC/Knowledge Availability Center. It is described as publishing two machine readable data bases, and no printed associates to these data bases. Special services include seminars and workshops by arrangement through the marketing department, machine and manual searching, selective dissemination of information (SDI, or "current awareness" searching), analytical reports, and technical specialists to clarify, summarize, and analyze the results of a search. This list can already be seen to be incomplete, for Information Market Place does not describe NASA as offering retrospective searching. "Retrospective searching" seems to have been a standard category in Information Market Place's questionnaire, so it should have been listed. This service is described as a NASA service in Spinoff 1979[10]

Of the twenty-six data base publishers, there were private profit and non-profit organizations, governmental agencies, and the United Nations. Not enough was known about the financing of these publishers to determine their relative numbers, but it seemed that the private organizations were the majority.

Table 8 illustrates the distribution of data base publishers by the number of machine readable data bases they produce:

TABLE 8

DISTRIBUTION OF DATA BASE PUBLISHERS

By Number of Machine Readable Data Bases Published

Number of Data Bases Published	Number of Publishers
One	7
Two	5
Three	8
Four - Nine	3
Ten or More	3

The three publishers producing more than ten computerized data bases are Cordura Publications Inc. with 26, Chemical Abstracts Service with 14, and the National Library of Medicine with 12. Twenty of the twenty-six data base publishers were responsible for three or fewer data bases. Thus, the industry seems to be characterized by a large

[10] Spinoff 1979, page 112

number of publishers who produce a small number of data bases, and by a small collection of publishers that produce a large number of data bases. Appendix F discusses a possible economic model to analyze this finding, and raises issues for further investigation.

The final table in this survey of data base publishers is a summary of data base related activities undertaken by the publishers. Table 9 lists the number of data base publishers that engage in the activity or service listed (numbers include the NASA facility). Those entries marked by an asterisk indicate NASA-IAC/Knowledge Availability Center activities, as reported in Information Market Place.

The first comment to be made on these results is that no definitions for the listed categories were provided. Thus, there may not necessarily be a clear distinction between certain of the categories. Furthermore, it is questionable that these descriptions are complete. As discussed previously, this data appears to have been completely self-reported by the respondents. Thus, there is reason to question the consistency of the responses. It is likely that if these activities were well defined, and if one interviewer had evaluated the activity of each publisher, then the numbers in some categories would be higher. For example, if 14 publishers engage in "machine searching" of data bases, it would seem likely that more than three engage in "retrospective searches."

TABLE 9

NUMBER OF DATA BASE PUBLISHERS

Engaging in Listed Activity

Activity	Number of Data Base Publishers
Publishes Print Products (often a paper version of the machine readable data base)	20
Training Programs	4
* Seminars and Workshops	14
User's Guide	6
Newsletter	4
* Machine Searching	14
* Manual Searching	8
Retrospective Searching	3
* Selective Dissemination of Information	9
Document Delivery	8
Facsimile Service	2
* Analytical Reports	5
Indexing	10
Telephone Interviewing	4
Thesaurus for Indexing	2
On Line Document/Hardcopy	1
* Access to Technical Specialists	5
Technical Conference Support	1

However, despite the potential problems cited above, one can note three items of interest. First, in providing "technical specialists to analyze and sum results, and to clarify questions," NASA is one of only five publishers listed that offer this type of customized service.

Second, only the ARPANET Network Information Center provided on-line document/hardcopy reproduction. Eight publishers (not including NASA) provide document delivery--such as the Institute for Scientific Information.

Except for these efforts, however, on-line access to documents remains a rare service.

Third, the NASA IAC/Knowledge Availability Center and the Petroleum Information Corporation were the only publishers to offer both analytical reports and access to technical specialists. This indicates that NASA's coupling of analysis by technical specialists with its data base is a relatively unique service in the data base industry.

6. DISTRIBUTORS OF MACHINE READABLE DATA BASES

This study was not oriented towards the analysis of machine readable data base distributors, except insofar as previous discussions have addressed this section of the data base industry. Historically, the IAC's have been the access point to RECON/NASA for all non-NASA users. With dial-up access to RECON being provided for more and more users, the role of the IAC's could change. It is important to clarify who will have dial-up access to RECON (it may never be unviversal), and to what extent this will change the nature and extent of the IAC's "clientele."

A parallel question is what the effects on the IAC's role would be if a machine readable version of the Tech Briefs were offered through a distributional vendor such as Lockheed, SDC, or BRS. An analysis of these effects is conditional, of course, on the extent to which one believes there would be a commercial market for this information. If there were little market potential, then no commercial vendor would offer such a file through their system (barring subsidies).

Although it is not the intent of this study to investigate future strategies for the IAC's, it does seem important to recognize that certain developments in the access to RECON/NASA's information could impact the role played by these institutions.

7. CONCLUSIONS

1. Technical and scientific data bases cover a wide range of material through a wide variety of formats, scope, and sizes. Of note is the finding that bibliographic data bases tend to cover multiple subject areas, whereas non-bibliographic ones tend to have single field coverage.

2. RECON/NASA contains some information of minimal use to those outside NASA. However, for that information that is of non-NASA interest, NASA data bases are much like other scientific and technical data bases in their structure, but historically have been much different in terms of their access.

3. Despite some user aggravation in certain types of searches, the structure of data bases seems tailored to user needs. Ongoing refinements and the development of new types of data bases should even more carefully tailor data bases to user needs.

4. The data base industry has a few large publishers of many data bases, and many small publishers. This observation carries with it as yet unanswered questions.

5. In terms of the activities of data base publishers, NASA seems well specialized in its provision of

technical staff for analytical and consultative purposes.

6. With increasing dial-up access to RECON/NASA, there may be reason for re-evaluating the role of the IAC's. If this is done, it is important to emphasize their current specialization in analytical and consultative services.

Appendix A

CONTENTS

Source: Information Market Place: 1978-79

Appendix B

Subject area headings from Information Market Place, 1978-1979 used to select data bases of direct relevance to a private sector scientist or technologist:

Aerospace and Aeronautical Engineering
Agriculture and Agricultural Engineering
Astronomy
Biology
Chemistry and Chemical Engineering
Civil Engineering
Computers, Data Processing Systems
Current Research Projects
Earth and Space
Electronics and Electrical Engineering
Energy
Engineering
Environment
Food Science
General Science and Technology
Geology
Life Sciences
Mathematics
Mechanical Engineering
Medicine
Metallurgy
Nuclear Science
Patents
Petroleum
Physics
Pollution
Technology

Appendix C

The data bases selected for the sample studied in this report, and their publishers, are:

DATA BASE NAME	DATA BASE PUBLISHER
1. API Tech Index	American Petroleum Institute--CAIS
2. ASTM Infrared Data Base	Sadtler Research Laboratories, Inc.
3. Agricola	National Agricultural Library
4. Biosis Previews	BioSciences Information Service (BIOSIS)
5. CA Condensates (CA Con)	Chemical Abstracts Service
6. CA Subject Index	Chemical Abstracts Service
7. Cancerproj	International Cancer Research Data Bank: Smithsonian Science Information Exchange
8. Chemical Abstracts Service Source Index	Chemical Abstracts Service
9. Chemical Industry News (CIN)	Chemical Abstracts Service
10. Clinprot	International Cancer Research Data Bank Program (ICRDG)

11.	Computer and Information Systems Abstracts; Electronics and Communications Abstracts	Cambridge Scientific Abstracts, Inc.
12.	Conference Papers Index	(CPI)
13.	Defense Market Measures System	Frost and Sullivan, Inc.
14.	Discontinued Diode	Cordura Publications, Inc.
15.	Discontinued Thyristor	Cordura Publications, Inc.
16.	Drug Product Information File	American Society of Hospital Pharmacists
17.	Elastomers	Cordura Publications, Inc.
18.	Energy Conservation	Energy and Environmental Response Center
19.	EnergyLine	Environment Information Center (EIC), Inc.
20.	Environmental Impacts	Energy and Environment Response Center
21.	Food and Agricultural Chemistry	Chemical Abstracts Service
22.	Geological Reference File (GeoRef)	American Geological Institute
23.	Hydrological Information Storage and Retrieval System (HISARS)	Biological and Agricultural Engineering
24.	IRIS Infrared Information System	Sadtler Research Laboratories, Inc.

25.	Interface Integrated Circuits	Cordura Publications, Inc.
26.	Linear Integrated Circuits	Cordura Publications, Inc.
27.	Maritime Research Information Service	MRIS
28.	Medical Subject Headings (MeSH) Vocabulary File	National Library of Medicine
29.	Microwave Tubes	Cordura Publications, Inc.
30.	NASA RECON	NASA IAC/Knowledge Availability Center
31.	Optoelectronics	Cordura Publications, Inc.
32.	Pharmaceutical News Index (PNI)	Data Courier, Inc.
33.	Polymer Science and Technology	Chemical Abstracts Service
34.	Production Standards Format	Petroleum Information Corporation
35.	Requests for Comments (RFC's)	ARPANET Network Information Center
36.	Science Citation Index (SCI)	Institute for Scientific Information
37.	Small Business Data File	International Data Corporation
38.	TTD Keyterm Index	Institute of Textile Technology
39.	Total Marketing Analysis Research Service	DMS Inc.
40.	Transdex	Bell and Howell Micro Photo Division

41. Well History Control System (WHCS)	Petroleum Information Corporation
42. World Energy Supplies System (Worldenergy)	United Nations Statistical Office

Appendix D

DIALOG DATABASES - NUMERICAL LISTING

```
? FFILES
Accessible files:
    1* ERIC 66-79/SEPT
    2  CA SEARCH 67-71
    3  CA SEARCH 72-76
    4  CA SEARCH 77-79/VOL 91(16)
    5  BIOSIS PREVIEWS 76-79/OCT
    6  NTIS 64-79/ISS21
    7  SOCIAL SCISEARCH 72-79/WK36
    8  COMPENDEX 70-79/SEP
    9  AIM/ARM 67-76
   10  AGRICOLA 79/JUL
   11  PsYCH ABS 67-79/SEP
   12  INSPEC 69-77
   13  INSPEC 78-79/ISS18
   14  INSPEC-MECH ENGR 73-79/OCT
   15  ABI/INFORM 71>79/SEP
   16  PROMT 72-79/OCT
   17  FTS PRELALERT OCT 29
   18  F & S INDEXES 74-79/OCT
   19  CHEM IND NOTES 76-79/ISS46
   20  FEDEFAL INDEX 76-79/AUG
   21  (*offline*)
   22  EIS PLANTS MAY79 (TYPES $0.50 EACH)
   23  CLAIMS/CHEM 1950-1970
   24  CLAIMS/U.S. PAT 1971-1977
   25  CLAIMS/U.S.PAT ABS 78-79/AUG: SEE FILES 23,24,125
   26  FOUNDATION DIRECTORY 1979 ED.
   27  FOUNDATION GRANTS 73-79/APR
   28  OCEANIC ABS 64-79
   29  MET/GEOASTRO ABS 70-79/FEB
   30  (*offline*)
   31  CHEMNAME(TM) FILE
   32  METADEX 66-79/JUN
   33  WORLD ALUMINUM ABS 68-79/JUL
   34  SCISEARCH 78-79/WK36
   35  CONF DISSERT ABS 1861-1979/SEP
   36  LANGUAGE ABS 73-78/ISS66
   37  SOCIOLOGICAL ABS 63-79/IS801
   38  AMERICA: HIST & LIFE 63-78/IS803
   39  HISTORICAL ABS 73-78/IS803
   40  ENVIROLINE 71-79/AUG
   41  POLLUTION ABS 70-79/JUL
   42  PHARM NEWS INDEX 74-79/SEPT
   43  CA PATENT CONCORDANCE 72-78
   44  AQUATIC SCI ABS 78-79/APR
   45  APTIC 68-78/OCT
   46  NICEM 1977 ED.
   47  MAGAZINE INDEX 77-79/OCT
   48  PIRA 75-79/SEPT
   49  PAIS INTERNATIONAL 76-79/JUL
   50  CAB ABS 72-79/JUL
   51  FSTA 69-79/AUG
   52  (*offline*)
   53  (*offline*)
   54  ECER/EXCEP CHILD 66-79/MAY
   55  BIOSIS PREVIEWS 69-73
   56  ART MODERN 74-78
   57  PHILOSOPHER'S INDEX 48-79/MAY
   58  GEOARCHIVE 74-79/MAR
   59  FROST & SULLIVAN DB2 75-78/IS803
   60  USDA/CRIS 75-79/JUN
   61  LISA 69-79/FEB
   62  SPIN 75-79/AUG
   63  NAIS ABSTRACTS 78-79/JUNE
   64  CHILD ABUSE AND NEGLECT SEPTEMBER 1979 ED.
   65  SSIE CURRENT RESEARCH 78-79/JUL
   66  GPO MONTHLY CATALOG JUL76-79/SEPT
   67  WORLD TEXTILES 78-79/JUL
   68  EPB 74-79/JUL
   69  ENERGYLINE 71-79/SEP
   70  NICSEM/NIMIS 1978 ED.
   71  MLA Intl. Bibliography 76-77
   72  EXCERPTA MEDICA 75-79/IS821
   73  EXCERPTA MEDICA IN-PROCESS 79/IS839
   74  IPA 78-79/JUN
   75  MONT CONTENTS 74-79/SEP
   76  (*offline*)
   77  CONFERENCE PAPERS INDEX 73-79/AUG
   78  NATIONAL FOUNDATIONS 1979 ED.
   79  FOODS ADLIBRA 74-79/SEPT
   80  (*offline*)
   81  FTS US STAT ABS 71-79
   82  FTS US ANL TIME SERIES JUN79
   83  (*offline*)
   84  (*offline*)
   85  (*offline*)
   86  FTS INTL STAT ABS 71-79/AUG
   87  FTS FRN ANL TIME SERIES SEP79
   88  (*offline*)
   89  (*offline*)
   90  ECONOMIC ABSTRACTS INTL 73-79/SEPT
   91  (*offline*)
   92  EIS NONMANUFACTURING MAY79 (TYPES $0.50 EACH)
   93  U.S. Political Science Documents 75-77
   94  SCISEARCH 74-77
   95  HAPRA ABSTRACTS 72-79/SEP
   96  (*offline*)
   97  RILM 6/20/79
   98  F & S INDEXES 72-75
   99  WELDASEARCH 1976-79/JUN
  100  DISCLOSURE 78-79/WK43
  101  (*offline*)
  102  (*offline*)
  103  (*offline*)
  104  (*offline*)
  105  FOREIGN TRADERS INDEX 79/OCT
  106  TRADE OPPORTUNITIES/77-7908
  107  TRADE OPPORTUNITIES/791030
  110  AGRICOLA 70-78/DEC
  111  NATIONAL NEWSPAPER INDEX 79-79/OCT
  112  Aquaculture 79
  113  (*offline*)
  114  ENCYCLOPEDIA OF ASSOCIATIONS ED. 13
  124  CLAIMS/CLASS JUNE 1979 ED.
  125  CLAIMS/U.S. PATS.ABS. WEEKLY 10/02/79
  200  DIALOG PUBLICATIONS 79/SEP
  201  ONTAP ERIC
  204  ONTAP CA SEARCH 15,785 DOCUMENTS
  211  (*offline*)
  221  ONTAP CHEMNAME(TM) 26,696 SUBSTANCES
  911  MEMSEARCH
```

*Trademark Reg. U.S. Pat. & Trademark Office.

Source: Guide to Dialog Searching, Lockheed Corporation

Appendix E

The publishers of the sample data bases, listed alphabetically:

1. American Geological Institute
2. American Petroleum Institute
3. American Society of Hospital Pharmacists
4. ARPANET Network Information Center
5. Bell and Howell Micro Photo Division
6. Biological and Agricultural Engineering
7. Biosciences Information Service (BIOSIS)
8. Cambridge Scientific Abstracts, Inc.
9. Chemical Abstracts Service
10. Cordura Publications, Inc.
11. DMS, Inc.
12. Data Courier Inc.
13. Energy and Environment Response Center
14. Energy Information Center
15. Frost and Sullivan Inc.
16. Institute for Scientific Information
17. Institute of Textile Technology
18. International Cancer Research Data Bank Program
19. International Data Corporation
20. Maritime Research Information Service

21. NASA IAC/Knowledge Availability Center

22. National Agricultural Library

23. National Library of Medicine

24. Petroleum Information Corporation

25. Sadtler Research Laboratories Inc.

26. United Nations Statistical Office

Appendix F

A Theoretical Discussion of the Data Base Industry Structure

An attempt was made to perform some theoretical analysis of the data base industry structure. This analysis was inspired by the works of Baumol, Fischer, and Braunstein, who have investigated cost and revenue functions from the perspective of firms producing heterogeneous goods. This perspective emphasizes a multi-dimensional analysis of the revenue and production costs of every combination of produced goods, by looking at revenue and cost behavior along and between many "output rays" in "production space." The key point is that the production of certain combinations of industrial goods may be characterized by complementarities of production and/or complementarities in consumption. If there are such complementarities, then the industry's production costs will decrease and/or the revenues will increase, respectively, with multi-good production. Thus, the industry will find it more profitable to produce particular combinations of several kinds of goods. Conversely, if a combination of goods is subject to

substitution in production and/or consumption, then the industry will be characterized by firms each producing a specialized and unique product.

If it is possible to demonstrate that there are complementarities or substitutions in production and/or consumption of certain goods (here--types of data bases), then one could demonstrate that there may be "natural" limits and "ranges of production" for firms in an industry. This would be counter to historical Economic wisdom, which has often labelled an industry with a few large producers and many small ones as "oligopolistic" and hence inefficient. Does the publishing activity of the sample publishers indicate a data base industry oligopoly, or a reflection of natural consumption and production forces?

This author, much to his regret, has no well-formulated answer to the above question.

There is great appeal in hypotheses focusing on the consumption side, and how this could largely determine a firm's size of production. Chemical Abstract's data bases most likely address much different user needs than do Cordura's or Sadtler Research Labs'. The six Chemical Abstracts data bases in the sample are bibliographic, whereas Cordura's seven and Sadtler's two are all non-bibliographic. The temptation is to assume that Chemical Abstracts data bases address users who would have

"complementarities" in consumption. This would indicate that Chemical Abstracts should produce more data bases and types of data bases than the other two. This distinction, however, does little to illustrate why both Chemical Abstracts and Cordura are "large" (14 and 26 data bases, respectively), and Sadtler is "small" (3 data bases). A proper evaluation of this issue would involve investigating all data bases produced by each of the twenty-six publishers. However, on the basis of the data bases included in the sample, and at this level of analysis, this does not appear a fruitful avenue of further research.

Looking to the production side for possible complementarity or subsitution of production, there are some a priori reasons to expect either substitutions or complementarities in the production of many types of data bases. The least controversial reason to expect complementarity in production is that producing a data base involves extensive organization and the writing of general software--each of which is a major front-end investment. Once made, these investments can be applied to the production of other data bases. Furthermore, if a firm owns its own computing facilities, then it is usually in the company's interest to spread the high fixed costs of this equipment over as much work as possible. Validating this would require examining the production processes and facilities of each data base publisher.

On the other hand, it could be possible that producing certain types of data bases involves exceptionally high production costs for a very specialized product, thus creating the potential for substitutions in production. For example, suppose there are extremely high fixed costs in developing data files and software to store and access chemical information by molecular formula. This could be an example where it is economic for a firm to produce only one or a few data bases, since the technology involved in these data bases may not easily be applied to other data bases. Yet, such data file and software technology would probably find success in separate data bases covering topics such as physical chemistry, organic chemistry, biochemistry, general medicine, pharmacology, toxicology, materials science, etc. Perhaps a better example would be to determine if Sadtler Research Labs has such a specialized expertise, and such specialized technological needs for its infrared spectral information data bases, that there would be diseconomies for Sadtler to produce other types of data bases.

An empirical study of cost functions would be of use in investigating this question.

There is an empirical study of the journal publishing industry which is interesting to cite in relation to this discussion. Baumol and Braunstein have performed an empirical study of scale economies and production

complementarities for a sample of nonprofit publishers of scientific journals.[11] Baumol and Fischer estimated a variety of cost functions, and found that about two/thirds of the firms were close to their "minimum cost locus," so that neither merging nor splitting these firms would reduce their costs. Furthermore, up to the "minimum cost locus," or "point of minimum ray average cost," costs per journal declined with the number of journals per publisher. The remaining one/third of the publishers was said to be in the region of "sub-additivity," such that limited amalgamation might reduce the firm's cost. As the journal publishing industry is more established than the relatively new machine readable data base industry, it would be interesting to see how close machine readable data base publishers are to the points of "minimum ray average cost."

---

[11] Baumol, W. and Braunstein, Y.: "Empirical Study of Scale Economies and Production Complementarity: The Case of Journal Publication," Journal of Political Economy Volume 85, number 5; October 1977; pages 1037-1048

BIBLIOGRAPHY

1. American Society for Information Science; Williams, Martha E.: Computer Readable Data Bases: A Directory and Data Sourcebook Washington, D.C.; 1979

2. Baumol, William, and Braunstein, Yale: "Empirical Study of Scale Economies and Production Complementarity: The Case of Journal Publication;" Journal of Political Economy volume 85, number 5; October, 1977; pages 1037-1048

3. Baumol, William, and Fischer, Dietrich: "Cost Minimizing Number of Firms and Determination of Industry Structure;" Quarterly Journal of Economics volume XCII, number 3; August, 1978; pages 439-468

4. Christian, Roger W.: The Electronic Library: Bibliographic Data Bases 1975-76 Knowledge Industry Publications; White Plains, New York; 1975

5. CLASS On-Line Reference Services: California Resources for On-Line Reference San Jose, California; 1979

6. Cuadra Associates (Editors: Ruth Landau, Judith Wanger, Mary Berger): Excerpts from Directory of Online Data Bases Santa Monica, California

7. Doszkocs, Tamas E.; Rapp, Barbara A.; Schoolman, Harold M.: "Automated Information Retrieval in Science and Technology;" Science volume 208, number 4439; American Association for the Advancement of Science; April 4, 1980; pages 25-30

8. EUSIDIC; Senders, James B.: Information Market Place 1978-1979: An International Directory of Information Products and Services R.R. Bowker; New York, New York; 1978

9. Information Alternative: Directory of Fee Based Information Services, 1977 Woodstock, New York; 1977

10. Information Industry Association: Information Sources 1979-80: The Membership Directory of the Information Industry Association Washington, D.C.

11. Lockheed DIALOG Information Retrieval Service: Guide to DIALOG Searching Palo Alto, California; November, 1979

12. Office of Space and Terrestial Applications: Technology Transfer Office: Spinoff 1979, and Spinoff 1980 Washington, D.C.; February, 1979, and April, 1980