16.127
August 17, 1965
Laboratory Report
IBM Confidential

SCIENTIFIC INFORMATION STORAGE AND RETRIEVAL: A SURVEY OF THE PRINCIPAL TYPES OF SYSTEMS

F. B. Wood

International Business Machines Corporation
ASDD Los Gatos Laboratory
Los Gatos, California

ABSTRACT

This survey examines examples of well-established systems, advanced systems and proposed systems for storing and retrieving scientific information, and extends into other areas with overlapping interests. Some systems for managing national economic data are included as particularly instructive examples. In the sciences, the varying approaches all aim to eliminate unnecessary duplication, to shorten the delay between the discovery and the dissemination of new knowledge, to respond more directly to the individual user, and to ensure workable connections with industry, the universities, and government and other agencies. These goals call for the efforts of systems engineers, electronic engineers, behavioral scientists, communications experts, librarians and other information specialists. Despite some interest in a centralized national information service for the sciences, various considerations--largely technological and sociological--seem to recommend decentralized, discipline-oriented systems connected by a national network.

LOCATOR TERMS FOR THE IBM SUBJECT INDEX

Information, scientific
13 Management Sciences
23 Miscellaneous
Storage, information
Retrieval, information

CONTENTS

PREFACE

To give the Advanced Technology Department a better perspective for its activities in information retrieval, Dr. F. B. Wood was asked to prepare a report on the present status of scientific information storage and retrieval systems. Dr. Wood has done an outstanding job, giving the reader a feeling for a wide variety of information retrieval activities, without creating confusion. This report can serve as a guideline--showing how IBM can help resolve the information management problem by providing the necessary technology and overall systems management (the latter being nonexistent in most systems).

Urgent needs have forced government organizations and private companies to tackle their own problems and find temporary solutions. There is yet to be found a constructive way to assure a brighter future for information management, and to prevent a crisis which could be a great impediment to progress. There are real opportunities for IBM to help solve urgent national and international information retrieval problems by filling recognizable market needs.

Some of these needs can be noted as one reads Dr. Wood's report. One concludes, for example, that very large files--random access and sequential, updatable and suitable for high-speed search--are not on the market but are needed in numerous information centers to make efficient automated searching possible. The Library of Congress represents one of the most outstanding examples of this need.

Another great impediment to progress is the lack of low-cost message communications for inquiry-reply traffic to and from information centers. The third but not least urgent need is for an overall systems approach to assure systematic procedures and orderly progress toward a successful solution of the information management problem on a national and international level.

Emil Hopner, Manager
Advanced Technology
ASDD Los Gatos Laboratory

I. INTRODUCTION

Scientific information retrieval is recognized as a national concern (see the Weinberg report to the President, "Science, Government and Information,"[1]), and as an imminent opportunity for those prepared to design and implement advanced information-handling systems. To appraise this opportunity realistically, we must understand the advantages and limitations of the present and projected systems for collecting, organizing and distributing scientific information. To this end, we will (1) survey the specialized information centers now serving the scientific community and examine several examples in detail; (2) describe the more notable advanced systems already in operation, and some methodologies being explored in small-scale pilot systems; and finally (3) outline the major proposals for future systems.

Scientific research depends on outside sources for funding and for specific programs, and may result in changes (e.g., new products) which affect the national economy, military policy, etc.; similarly, scientists rely on outside sources of information and in turn generate information vital to many extra-scientific interests. For this reason, as Weinberg pointed out in his "Second Thoughts

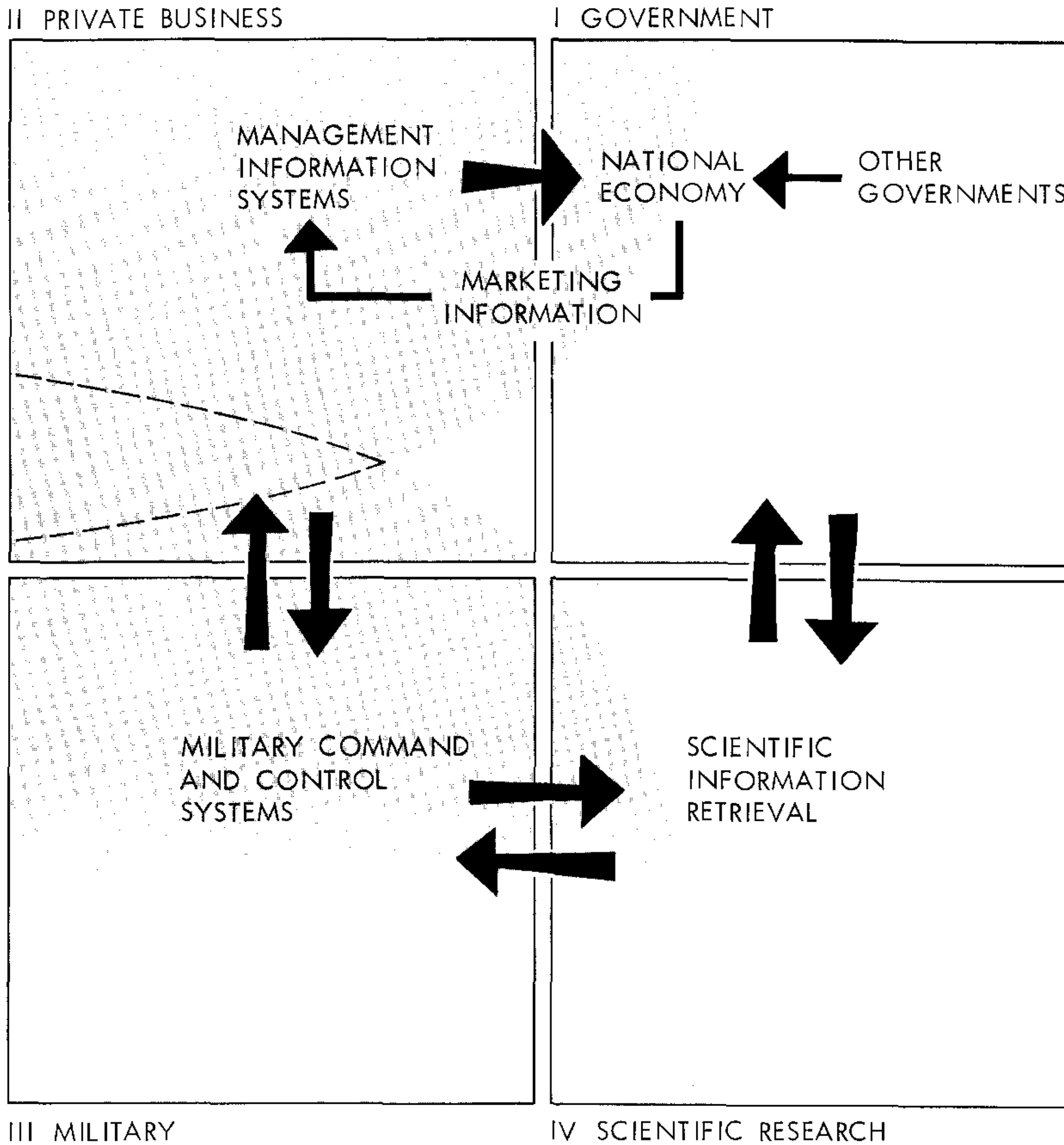

Fig. 1. Information Management in the United States (Shaded areas represent classified information.)

on Scientific Information Retrieval,"[2] the problem is not an isolated one, but has general implications for the social organization of science. Scientific information retrieval must be viewed as part of the larger problem of information management, which can be described roughly as in Fig. 1. The dividing lines there are arbitrary, but the approximate boundaries of the problem can be seen, and the overlapping interests appear in something like their real relationships.

In Fig. 1, the first quadrant (I) represents the civilian branches of government; quadrant II, the business sector; quadrant III, the military; quadrant IV, scientific research. Quadrants II and III involve primarily private information, i.e., proprietary business information and confidential military information, respectively. The other two quadrants involve chiefly what we might call public information (although some segments coincide with classified information). Our main interest here is in the unclassified portion of scientific information--the facts which are in the public domain and can be disseminated freely. The problem is how to disseminate them more effectively. As we will see, some information management systems designed for other quadrants prove relevant to the problems of scientific information, and some national networks established or proposed elsewhere (in France and in the U.S.S.R.) suggest possible ways of managing information in accordance with national purposes.

We can get some idea of the scope and magnitude of the scientific information problem by looking at recent activity in a few specific areas. In the physical and biological sciences alone, according to a 1961 survey by the National Science Foundation,[3] more than 400 science information services had sufficient staff and sufficiently comprehensive collections for literature searching in their special subjects. The 427 agencies serve twenty-four different branches of science as listed in Table 1.

Although each area has some relatively independent sources of information, the branches of science are interdependent in many ways. Each needs ready access to the others and to information not primarily scientific. Research in psychiatry, for example, as one of the medical sciences, could reach into literature (see Ref. 4 for a sample list of related drama, myths, fairy tales, poetry, scripture). An ideal system would make such cross-references as easily accessible as the primary references.

TABLE 1
DISTRIBUTION OF INFORMATION CENTERS BY NUMBER PER BRANCH OF SCIENCE

Centers per Branch	Branches of Science
1 to 10	Documentation and Graphics Food and Nutrition Human Engineering and Management Materials-Nonmetallic (Inorganic) Mathematics Ordnance Pharmacology
10 to 20	Aerospace Agriculture Architecture-Civil Engineering Combustion, Fuels Electronics, Electrical Engineering Health Mechanical Engineering Nuclear Physics Physics General Science
20 to 30	Chemistry Conservation Materials-Metallic Organic Materials (not including petroleum)
30 to 40	Medical Sciences
40 to 50	Biological Sciences Earth Sciences

The size of a collection of specialized scientific information affects both its usefulness and the feasibility of automation. As more data is published on each subject, manual methods become inadequate. The National Science Foundation identified a number of advanced science information centers--services employing some automated methods of storage and retrieval, and retrieving information from as many as one million documents. Of 56 services active in the years 1959-1961, 23 had fewer than 10,000 documents each; 15 had 10,000 to 25,000; 12 had between 25,000 and 100,000. Three were in the next range, 100,000 to 200,000, and 3 others had 200,000 to 1,000,000 documents each.

Since these studies, one important change has been the development of a number of "data banks" in the social sciences. These include archives which are "limited to coded data accumulated in machine-readable format...not simply for storage, but primarily to facilitate analysis."[5]

II. SPECIALIZED SCIENCE INFORMATION SERVICES

In this country and abroad, the older specialized science information services are generally oriented toward well-established disciplines or professions; newer services often concentrate on a single emerging phase of a science. Nuclear physics, for example, is covered intensively by one of the specialized centers described below. With increasing government responsibility for scientific research (medical, military and other), new and reorganized information services sponsored by the government are developing into large-scale systems, as will be shown.

A. SERVICES IN THE UNITED STATES

1. ENGINEERING INDEX (New York)

Started in 1884, this is the oldest abstracting service of substantial size in this country and covers all fields of engineering and related physical science.[6] The abstracting staff scans more than 1400 journals, indexing only the more significant articles.

A conference is usually entered as one abstract which lists the authors and titles of the papers presented there. The number of abstracts per year has risen from 36, 000 in 1960 to 51, 000 in 1964. The abstracts, classified into 249 fields of interest, are issued weekly on 3 by 5-inch cards. Customers can subscribe to complete sets of weekly cards for $1500 a year or partial sets (for particular fields) at proportionate rates. Monthly and annual indexes containing all abstracts issued during the period are printed for $250 per year ($75 for annual volume only).

The Engineering Index does not provide other services, but its depository, the Engineering Societies Library, performs literature searches ($6 per hour); supplies translations (quotations on request); retains copies of all material abstracted and supplies photocopies of papers not available to readers locally.

Engineering Index has been working with the Engineers' Joint Council on a computer system for mechanized indexing. [7] The Tripartite Committee, representing these two groups and United Engineering Trustees, is preparing the organizational framework for an integrated information system for engineering.

2. CHEMICAL ABSTRACTS (Columbus, Ohio and Washington, D. C.)

This service, begun in 1907, provides abstracts of all new chemical and chemical engineering information, in publications prepared at Ohio State University for the American Chemical Society. The sources are journals, government reports, industrial literature, and patents, American and other.

The basic Chemical Abstracts monitors 10, 000 publications, and each bi-weekly issue contains about 8000 abstracts. The annual number of signed abstracts has increased from 133, 000 in 1961 to 200, 000 in 1964. These entries are classified under 74 main subjects, and indexed by keywords and by author. Patents are indexed separately by number, and a patent concordance relates corresponding patents of different countries. This service costs $1000 per year, or half that for colleges and ACS members. The collective index, issued every five years, costs $1400 and includes a formula index.

A supplementary semi-monthly listing, Chemical Titles ($50 per year), gives the tables of contents of 650 major chemical journals. Permuted titles form a keyword index (KWIC) prepared on an IBM 1401, and an author index is included. An additional publication, the Ring Index, lists the chemical rings reported in the world's chemical literature (11, 524 to date). Special literature searches are performed on request, prices by quotation.

3. NUCLEAR DATA PROJECT (United States Atomic Energy Commission, Washington, D. C.)

This center collects experimental data on nuclear energy levels, related information in basic nuclear physics, and takes the responsibility for correlating all reported data on nuclear structures. The source material includes technical journals, abstract journals, government reports, and technical correspondence originating in the United States and elsewhere. Data on a given subject is organized to keep a running account of experiments and results reported. The data is summarized in a form which permits the reader to survey quickly the sequence of developments and the conclusions to date.

Table 2, for example, covers certain data on europium-152 (Eu-152).[8] As of December 1963, 164 papers had been published on this subject, and new information is reported in an average of two papers per month. The Nuclear Data Project digests all these papers, analyzes the new data, and prepares tables listing the significant results reported. Table 2 summarizes the findings of experimenters measuring the lowest gamma energy in the decay of 13-year Eu-152. As the measurement becomes more precise, the observed values of γ_1 appear to converge at 121.78 kev.

This example shows how useful a scientific information service can be, and how directly that usefulness depends on a staff of competent specialists who can evaluate the accuracy, quality and significance of information. In the case of NDP, the staff prepares replies to technical questions from individuals in addition to surveying all pertinent publications.

As summaries are updated, the NDP also issues Nuclear Data Tables, Nuclear Theory Index Cards, Nuclear Data Sheets, and lists of Radiations from Radioactive Atoms in Frequent Use.

TABLE 2
TYPICAL INFORMATION SUMMARY
PREPARED BY NUCLEAR DATA PROJECT

Experiments Reported*	γ_1 Energy (Kev) in Decay of 13-Year Eu-152
Shull (1948) [1]	123
Cork (1950) [2]	121.8
Fowler (1956) [3]	121.4
Grodzins (1956) [4]	122
Bobykin (1957) [5]	121.77 ± 0.1
Cork (1957) [6]	122.2 ± 0.2
Nathan (1957) [7]	122
Anderson (1958) [10]	121.87 ± 0.06
Romanov (1958) [11]	121.75 ± 0.03
Hatch (1959) [12]	121.79 ± 0.03
Mukherjee (1960) [14]	122.1
Marklurd (1963) [15]	121.778 ± 0.006

*Superscript reference numbers identify the source of information in each case.

4. THERMOPHYSICAL PROPERTIES RESEARCH CENTER (Purdue University, Lafayette, Indiana)

This center collects, classifies, codes, and disseminates reported information on a total of 30 thermophysical properties of matter, tabulating the data to indicate the most probable values of given properties of specific materials. The TPRC collection consists primarily of abstract journals and government reports, some of them classified; it includes all relevant journals, periodicals, books, industrial literature, and technical and scientific data available in other forms. Staff specialists (including experts in thermodynamics and information theory) search the literature and evaluate the accuracy, quality, and significance of the data found. Consulting services include conferences and contract research. For example, TPRC has undertaken research on information retrieval for the National Science Foundation.[9]

The cyclic searching system developed, which shows high yield at relatively low cost, first searches abstracts for "a" years by subject classification, keywords and/or titles. Then the citations in these Class A references are searched as Class B references. When the plotted yield of useful references per year falls off, at year "b," a new cycle for the period (b, b+a) years is searched in the abstract journals. The citations yield a set for the approximate period (b, 2b). In an alternative "cascade" method, TPRC searches by abstracts, keywords, etc., only for the first "a" years, thereafter following citations, citations in cited references, and so on. By means of a mathematical formulation developed, TPRC can compute yield and cost for these alternate searching methods.

The Center publishes these annual reports: Retrieval Guide to Thermophysical Properties Research Literature ($120) and Masters' Theses in Pure and Applied Sciences Accepted by Colleges and Universities ($6). The Data Book of "Most Probable Values" has been issued nonperiodically in three volumes, and looseleaf additions are distributed as they are completed.

5. NATIONAL LIBRARY OF MEDICINE (Washington, D.C.)

Now operated by the U.S. Public Health Service (Department of Health, Education and Welfare), this was formerly the Army

Medical Library and the Armed Forces Medical Library, and in its earliest form, the Library of the Surgeon General (established in 1836). In 1879, current medical literature was first indexed on a large scale by the institutional team which produced the first volume of Index Medicus. This monthly publication now contains references to more than 10,000 articles per issue, and an annual Cumulated Index Medicus is also published.

Since 1964, Index Medicus has been prepared automatically by the Medical Literature Analysis and Retrieval System (MEDLARS) which will be described in Section III. The National Library of Medicine also performs literature searches in response to specific requests. Such services are expedited by the availability of government communication channels -- the General Services Administration teletypewriter network, for example, with terminals in government hospitals, public health departments, etc.

6. CLEARINGHOUSE FOR FEDERAL SCIENTIFIC AND TECHNICAL INFORMATION (Springfield, Virginia)

As the central source for government research data in the physical sciences and engineering, this Clearinghouse replaces the Office of Technical Services (Department of Commerce). It reproduces and distributes all unclassified Department of Defense research reports to the public and to DOD agencies and contractors. Clearinghouse also keeps track of federally-sponsored research in progress, and of sources of technical expertise in the government. Data processing and reproduction equipment speed the flow of the information from Clearinghouse.

Twice a month, Clearinghouse issues a listing called U.S. Government Research and Development Reports (USGRDR) which announces the public availability of reports on government-sponsored research and development, and lists government-owned patents released for public use. One part of this publication, the "Technical Abstracts Bulletin," covers reports released by the Defense Documentation Center (DDC), and the second part lists reports released by civilian agencies. Bi-weekly Technical Translations bulletins list recent and scheduled translations of foreign technical reports. Nuclear science reports, AEC engineering drawings, and aerospace reports appear in separate abstract publications issued by the AEC and by NASA, but can be ordered from Clearinghouse.

7. CORPORATE INFORMATION RETRIEVAL SYSTEMS

In large corporations engaged in engineering research and development, centralized information retrieval systems typically contain internal reports, patent disclosures, publications released to the public, and reports received from external agencies. The more up-to-date systems maintain interest-profiles for all subscribers, so that computer searches on new entries can automatically prepare bibliographies to match user interests. The documents indexed may be kept available on microfilm or may be identified by number for retrieval as printed copies: internal reports by Company number, external by Clearinghouse number. Duplicate magnetic tape reels may be distributed for local searching at regional centers within the corporation. To request a search on a particular subject, the user can specify a set of keywords and their logical relationship. If a computer search of the local tape is not enough, a supplemental search can be undertaken at the company's main information center.

B. FOREIGN INFORMATION SERVICES

Two information services outside the United States are of particular interest to us--one because scientists and engineers in this country continue to rely on it; one because it is more extensive than any yet developed in this country. The first, Science Abstracts, has been published in London since 1898 by the Institution of Electrical Engineers in cooperation with other scientific societies. Monthly abstracts are issued in two parts: Section A: Physics (averaging 26,000 abstracts per year), and Section B: Electrical Engineering (averaging 16,300). The entries, organized by the Universal Decimal Classification system, are indexed by subject and author in each issue, and in an annual cumulation.

This service still provides the most thorough abstract coverage of general physics available to American physicists. In 1964, Science Abstracts summarized work in some 800 journals, more than 100 of these being abstracted completely. Additional abstracts represented papers prepared for approximately 115 conferences that year.

The second non-American service of special interest is the All-Union Institute of Scientific and Technical Information (VINITI), Moscow, U.S.S.R.[10] First organized in 1952 by the U.S.S.R. Academy of Sciences as the Institute for Scientific Information, VINITI now is responsible for keeping Soviet scientists informed about progress throughout the world. VINITI is part of a well-developed system which includes a large network of technical information organizations--84 at the national level, 94 regional bureaus, 4000 local bureaus, research institutes and production facilities, and more than 16,000 technical libraries.[11]

In VINITI, a permanent staff of 2200 publishes abstracts contributed by more than 20,000 reviewers; operates the Electromodeling Laboratory, the Laboratory for the Mechanization of Documentation, and the Department of Scientific Methodology in Information and Documentation.

VINITI publications organize the available scientific information in several ways: Referationyy Zhurnal (Abstracts Journal) consists of abstracts of world-wide scientific and technical literature, bibliographic citations of patents, monographs, and proceedings, and book notices and reviews. Ekspress Informatsiya selects non-Soviet research publications of special interest and presents them in the form of detailed abstracts of abbreviated translations. In Itogi Nauki (Scientific Achievements), retrospective studies are published to explain the state of the art in special subjects. In addition, bibliographic cards are prepared and disseminated to various organizations in the Soviet information network.

III. ADVANCED SERVICES IN OPERATION

Two kinds of advanced scientific information storage and retrieval systems are of special interest: complete systems which represent the latest technology and methodology, and pilot systems developed to test methodologies or to demonstrate feasibility.

A. OPERATING SYSTEMS

Among operating systems of substantial size, four can be taken as representative, showing various ways computers can help retrieve information, or prepare indices and bibliographies.

1. DEFENSE DOCUMENTATION CENTER (DDC) (Arlington, Virginia)

DDC provides a central service in the Department of Defense for the interchange of scientific and technical information. Both classified and unclassified documents are indexed in DDC, but the unclassified ones are now distributed through Clearinghouse. The Defense Documentation Center is the successor to the Armed Forces Technical Information Agency (ASTIA) and continues the AD report index. DDC started in 1963 and now includes more than 600, 000 documents.

Computer services make it possible to do these things automatically in the DDC system:

a. Validate need-to-know ("authority to ask") and check security clearance.

b. Control inventory and initiate document printing orders.

c. Account for all documents.

d. Prepare indices.

e. Check incoming documents for duplication.

f. Identify documents without catalog numbers.

g. Answer reference questions.

h. Prepare bibliographies and print out lists of report numbers or copies of bibliography cards. General bibliographies are prepared periodically; "demand" bibliographies in response to specific requests. For the latter, the computer searches on keywords within an indicated subject field.

2. MEDICAL LITERATURE ANALYSIS AND RETRIEVAL SYSTEM (MEDLARS) (Bethesda, Maryland)

In operation in the National Library of Medicine, MEDLARS is a relatively unsophisticated computer-based system which indexes reports and compiles abstracts. Coordinate indexing and Boolean searching are the principal techniques,[12] and the throughput of the system is considerable: it is currently processing 700 documents a day for the world's largest biomedical library. Now accessible by computer are 200,000 documents (this represents the number of accessions during 1964, when the present computer-based system began operating).

The system is designed for three major compile-and-print tasks, preparing (1) the Index Medicus, a monthly listing of medical journal references and abstracts (currently, 14,000 per issue), (2) recurring bibliographies, periodically updated, on cancer research, heart disease, mental health, etc. and (3) demand bibliographies. The major impetus for computer-based mechanization of the files came from interest in the latter.

In the present system, literature analysts classify the subject content of each article by assigning descriptors from the library's controlled list of medical subject headings (MeSH). The computer (Honeywell 800) processes the input paper tapes with the MEDLARS (magnetic) dictionary tape to produce two magnetic tape outputs: CCF (Compressed Citation File) for the retrieval sub-system, and PCF (Processed Citation File) for the publication sub-system. In the retrieval sub-system, requests for bibliographies are received by telephone, mail or teletype. A search specialist determines the proper keywords and the logic of the search, and the search request is punched into paper tape. Requests are batched and compared with each entry in the CCF tape, once or at frequent intervals if continuing interest is indicated in the subject.

In the publication sub-system, cards are punched specifying which sections of Index Medicus are to be prepared each day. The report selection cards are fed into the H-800 output module which searches the PCF tapes, sorts and edits the material to prepare another tape. This tape is the input to the Graphic Arts Composing Equipment (GRACE), an off-line photo-composer which prints 300 characters/second (from a font of 226 characters) on positive photographic film or paper nine inches wide. The film is automatically processed, then inspected and sent to the printer who prepares the plates.

In addition to its primary objectives, MEDLARS was designed to be open to improvement and extension. It might, for example, be expanded as a national, decentralized medical bibliographic system. Two possible methods are being compared: it might be desirable to (1) increase capacity and provide remote input/output facilities, or (2) distribute duplicate CCF tapes to regional centers.[13] It may also become necessary to store and retrieve graphic images of textual material, if the present manual microfilming system (Xerox Copy Flo and Xerox 914 Copier) becomes overloaded. As more journals are published and must be indexed, it may be necessary to automate the records of serials (periodicals) and prepare catalog cards for them on the GRACE equipment.

3. SCIENCE CITATION INDEX, INSTITUTE OF SCIENTIFIC INFORMATION (Philadelphia)

The Institute for Scientific Information maintains a computer-based file of citations of science articles. The basic file, alphabetic by author, lists citing authors and citing references for each source author. Every source item is identified by code as an abstract, letter, technical article, etc. The citations are published in quarterly print-outs, an annual cumulative index, and a source index with each volume (cost: $1950 per year, $1250 for colleges).

Such a system requires no staff of experts to select keywords and classes. This is the unique feature of citation indexing: the authors of published articles do the coding by their choice of citations, and therefore the system emphasizes not what has been published, but what publications have been found useful. The citation index is concept-oriented, i.e., organized on the basis of the relationships of concepts in the source and cited papers. The basic file is continuously updated and automatically re-indexed. Whenever a new article is listed, the lists for all its cited articles are updated. One can go either forward or backward from a given reference to trace the development of a special segment of science.

At present SCI covers only medical literature thoroughly; engineering and physical science are covered only in part, but this method makes very complete coverage possible. Whereas the leading medical abstracting service covered 110,000 documents in 1963 and 200,000 in 1964, SCI covered 1,000,000 in 1964 and expects to cover 2,000,000 in 1965. When SCI covers all current journals on a given subject, a negative search is conclusive.

An auxiliary service, Automatic Subject Citation Alert (ASCA) prepares individual print-outs for subscribers interested in specified subjects or authors. These print-outs are prepared automatically during the computer processing of the weekly additions to SCI, and cost the subscriber $2 per citation, $20 per author, or $100 per year.[14]

4. FRENCH ECONOMIC INPUT-OUTPUT TABLES

France has developed techniques for processing economic data in a large-scale information system with multiple sources and with cooperative, not coercive, controls. Government, industry, labor unions, teachers, and other social groups cooperate in collecting and analyzing pertinent data, and on the basis of this data, help draft plans for national economic development. This system interests us for several reasons: its success in keeping complex data current and usable (in the form of comprehensive input-output tables); its balanced sponsorship, governmental and private; its practical appeal to the enlightened self-interest of all participants in the national economy.

Taking a middle way between American and Russian approaches, French planning seeks to "indicate" what public and private groups can do to improve national productive efficiency and social welfare, without dictating what any business or person must do. Such planning has also been described as being not so much indicative as active;[15] i.e., many decision makers are associated with the preparation and revision of economic projections. The substantial work is done by 25 Commissions de Modernisation, 20 of them responsible for particular sectors of industry (iron, non-ferrous materials, chemicals). The other five have horizontal or cross-sectional responsibility for certain areas: labor, finances and general equilibrium, regional development, research, and productivity. Each Commission consists of 30 to 50 members from three general groups--public servants, business leaders, and labor leaders. These Commission members are joined by other interested citizens in study groups assigned to particular problems. Altogether approximately 3500 people participate in these activities, and their decisions and recommendations are carried out by the Commissaire Général du Plan Masse assisted by a small administrative staff (about 100 people).

In the input-output tables, the data gathered is "stored" in highly concentrated form for what might be called real-time analysis (in contrast to the usual delayed access to such data). These tables are prepared by the Department of Economic and Financial Research (in the Ministry of Finance) from statistics furnished by the National Institute of Statistics and Applied Economics (INSEE). The tables are calculated twice a year as the basis for short-term forecasts--in May, when the National Accounts Commission meets, and in October, when the appropriation bill is being drafted. These forecasts are based on figures for the prior year and relate to the current year and the next year so the input-output tables are adjusted accordingly.[16]

An input-output table prepared by Leontief for the United States in 1939 was similar except that it treated foreign trade separately; the French tables account for foreign trade within each industry sector. (Except for agriculture, the French have found that changes in technical coefficients--i.e., the relative positions of different sectors--have been due to specific technological changes in production.)

Guided by these widely available and well-publicized tables, French technical economic planners develop proposals in consultation with private and public interests. The method may be a promotional stimulus: the planners provide information on market capacity and this information raises incentives to invest or adapt production in new directions. Or--a second method--the planners may arrange for government-business consultations to encourage agreement on specific objectives and targets. The tables are useful in many ways: the technical coefficients in the input-output tables help market analysts in private firms estimate markets; bankers check the forecasts before approving loans for business expansion, etc.

B. PILOT SYSTEMS AND METHODOLOGY STUDIES

To test proposed methodologies and help specify the requirements for future mechanized library systems, some pilot systems have been designed and are in operation. SASIDS, for instance, is a computer-controlled system for disseminating abstracts scientists prepare for each other. Another pilot system, SMART, can store one set of documents under several different classification systems, and compare retrieval speed and accuracy. A third system is called the Library of the 21st Century, and is a pilot model of the services projected for future libraries.

1. SASIDS - STOCHASTIC ADAPTIVE SEQUENTIAL INFORMATION DISSEMINATION SYSTEM (NASA Project, University of California, Berkeley)

In this information exchange, 70 scientists and engineers prepare brief reviews of current articles they find valuable. These reviews are assembled and distributed weekly to all participants. The particular emphasis in SASIDS is not on the quantity of information, speed of dissemination, or degree of automation, but on relevance to the user: how to make sure the subjects covered match his changing interests. The quality of the information, i.e., the value to the user, is the test of the system, regardless of retrieval speed, storage capacity, etc. A successful system must be sensitive to both interest and disinterest on the part of the user.

To explore ways of keeping the system content relevant, participants both prepare and evaluate the abstracts. They mark-sense cards to indicate which reviews they found valuable, rating each on a scale from zero to 1.0. These evaluations provide a source table from which a computer program prepares a customer-reviewer correlation matrix. In this matrix, P(i/j) is the probability that scientist (j) is interested in the article that scientist (i) puts into the system. Being stochastic (i.e., it selects information on a probabilistic basis), the system generates a random number between 0 and 1 for each abstract sent in and compares that number to an 'adaptive' parameter to decide who is to receive the material. The scheme is adaptive in this sense: an initial parameter, $P_{ij} = 1$, is assigned as a link between member (i) and member (j). When an abstract is received from member (i), it is sent to (j) with probability P_{ij}. When (j) assigns a rating ($0 \leq R \leq 1$) to the abstract, parameter P_{ij} is modified by the formula $P_{ij} = u_j P_{ij} + (1 - u_j) R$. (Currently, $u_j = 0.8$ for all members.) The sequential nature of this process is evident.

This mutual interest parameter tends to stabilize between any two people while their interests remain stable, but is expected to shift as their relative interests change. The present goal is to maintain contact primarily in the areas of greatest interest, and to find ways to avoid distributing items of marginal or doubtful interest.

2. SMART (National Science Foundation project, Computation Laboratory, Harvard University)

The SMART system is designed to process English texts and search requests (i.e., natural language input) without prior manual analysis, and to identify for retrieval those items of information which are similar according to specified criteria. Programmed on an IBM 7094 and operating as a storage and retrieval system (since June 1964), SMART is also an instrument for evaluating retrieval systems. The system incorporates a large number of syntactic, statistical and semantic features, and is designed for repeated change and adjustment. In this it anticipates the man-machine environment in which the information requester will be permitted to keep modifying his search request until he is satisfied. In this sense, SMART can simulate the retrieval environment.[17]

To compare and evaluate different ways of handling the same set of documents, the system can vary the technique for determining information content; can change the criteria for matching items of stored information--or keep the same criteria but relax or tighten them; and can specify the request in different ways.

Under the control of a programmed supervisory system and chained together in FORTRAN II, the SMART subroutines can manipulate language data in several ways significant to the organization and retrieval of information. The system can convert original text into a series of hierarchical concept numbers, which in effect replace library or other file classifications. It can compute term correlation from co-occurrences within sentences, and compute term and document clusters from term co-occurrences within documents of a given collection. Further, the system can analyze syntax and compare texts by structure-matching procedures. These procedures are applicable to any problem whose data can be represented by graphs, whether the purpose is to match organic chemical compounds or syntactically analyzed sentences.

Four basic tables supply the notation by which information is identified. (a) The alphabetic stem dictionary supplies each word stem with a number of syntactic and semantic codes. (b) The alphabetic suffix table supplies syntactic codes for word suffixes. (c) The numeric concept hierarchy represents various relations between semantic categories. (d) The phrase dictionary, which consists of concept pairs rather than word pairs, aids syntactic processing. The many worthless word pairs (non-significant co-occurrences within a sentence) must be weeded out, and this is done by means of the Kuno Syntactic Analyzer and "criterion tree" (a dictionary of pre-analyzed syntactic phrases). Another source table may be added for bibliographic citation data, which is not now included.

Eight processing options are available for classification and searching. (a) General processing includes the input of new text, which is in effect a search request; i.e., it is matched with all the previously stored information. The other options are (b) alphabetic dictionary procedures, (c) concept hierarchy procedures, (d) statistical correlation of terms co-occurring within systems (or (e) within documents), (f) syntactic analysis by means of phrase dictionaries, (g) document matching, and (h) dictionary updating.

The performance of retrieval systems is measured in a number of ways (none yet standardized in general practice), but most commonly by "recall" and "precision." These reflect, respectively, "the proportion of relevant material retrieved, and the proportion of retrieved material relevant." (Such measurements depend, of course, on some systematic personal evaluations of relevance.) In a series of tests of the SMART system, nine different processing methods were applied singly and in combination to ten specific and seven general requests. Substantial differences were found in the results achieved with different methods, both recall and precision tending to vary in the same direction. The system performed better on the specific questions than on the general ones; i.e., performance was better for clearly specified logic classes. The most effective processes used combinations of concepts (i.e., phrases) rather than individual concepts alone.

Within the SMART project, the normalized overall effectiveness has been defined[18] as a weighted combination of recall and precision, having a maximum value of two. In the searches based on the general requests in the tests, the nine retrieval procedures ranged between 1.05 and 1.60. With the regular thesaurus procedure, the overall effectiveness of the system was 1.55; with word stems and statistical phrase runs added, this became 1.75. According to this measure, three different combinations of methods proved better than any one method alone;

such combinations appear to be the most promising search strategy.

3. LIBRARY OF THE 21ST CENTURY (Bolt, Beranek and Newman, Incorporated, Cambridge, Massachusetts; Sponsored by Council on Library Resources)[19]

In this pilot demonstration of a possible future library (see Fig. 2), new documents entering the system each day are processed by the "organizer," a staff of analysts aided by a powerful computing system. Each new document is (a) entered into a machine-readable "document store," (b) indexed (and the thesaurus and retrieval program updated accordingly); and (c) checked against the encyclopedia "information store" (which is also updated if necessary). As the figure suggests, a user can request copies of pertinent documents or can question the information store directly.

In the BBN pilot demonstration, the SYMBIONT program handles the document retrieval (coding to dispenser). This program runs on a PDP-1 computer, and the operator can enter and modify questions by typing control signals (on an electric typewriter) and by touching areas of the display screen with a light pen. QAS, the automatic question-answer system (programmed on an IBM 7090) shows its flexibility in the example below, where the same set of facts supplies correct answers to questions which vary considerably in form.

Is 3100 miles the diameter of Mercury?	Yes
Does Mercury have a diameter of 3100 miles?	Yes
What is the diameter of Mercury?	3100 miles
What planet has a diameter of 3100 miles?	Mercury
Does Mercury have a diameter of 3200 miles?	No
What planet has a diameter of 3200 miles?	No answer found
Is Saturn a planet?	Yes
What is the distance between Earth and Sun?	93 million miles
Is the moon a sun?	No
What is the period of revolution of Mars?	687 days
Do all planets have moons?	No
What type of celestial body is Andromeda?	I'm sorry, I don't understand the question

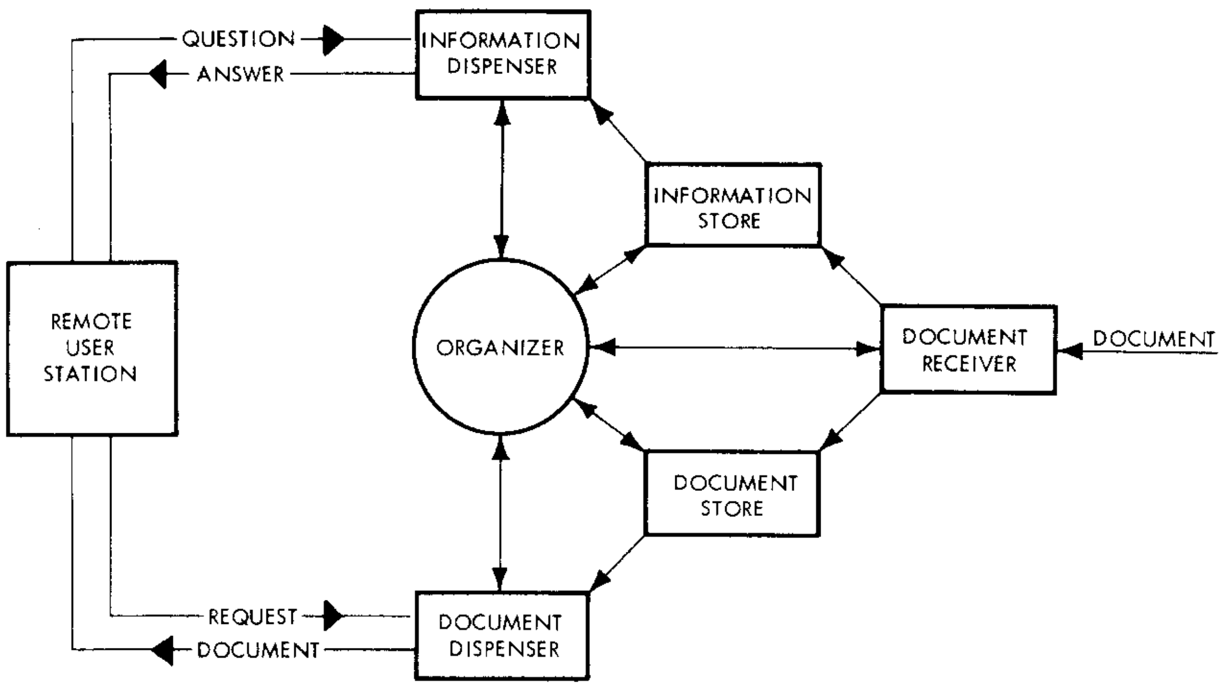

Fig. 2. Flow of Documents and Information in Possible Future Library

IV. PROPOSALS FOR FUTURE SCIENTIFIC INFORMATION STORAGE AND RETRIEVAL SYSTEMS

Of the numerous studies and systems suggested, we consider here only large-scale systems, definitely proposed, which would have a major impact either on information storage and retrieval in general, or on scientific information in particular.

A. REVISIONS OF PRESENT SYSTEMS

1. AUTOMATION OF THE LIBRARY OF CONGRESS

As a national depository of information, the Library of Congress seems a logical place for large-scale automation, since increased efficiency there would be of widespread benefit. The question has been not so much "Should it be done?" as "When should it be done?" A committee of the Council of Library Resources, considering both questions, recommended automation now, beginning with available technology. The committee report[20] pinpointed some of the problems that call for further technological development. For example, a fully automated catalog demands a trillion-bit memory, and this can be economical only with additional engineering development. Again, new ideas are needed in regard to practical display consoles; and systems research efforts must be undertaken to determine the optimum user-system relationship--what might be called the tutorial interaction between the two. The problem of creating the needed machine-readable files in the Library of Congress is complicated by the heterogeneous nature of some of the records involved. Optical sensing, one suggested method of file conversion, could not adequately handle the National Union Catalog, for instance.

According to the committee, automated bibliographical control is also technically feasible, and would make the Library of Congress more responsive to users' needs. An automated Library of Congress could materially speed up coordination of information services and could in fact become the control center for a national library system. The suggested automation, based on present technology, would be no more costly (through 1972) than the present manual system, and would be considerably more effective. These requirements were assumed in the committee report:

	Capacity (Bits)	Access Rate (Items/Minute)
Technical processing (ordering, receiving, etc.)	9×10^8	22
Cataloging (content classification)	5×10^{10}	170
Retrieval and circulation	1×10^9	39

Assuming a target completion date of 1972, the major functions (cataloging, category maintenance, acquisition, internal processing, circulation, retrieval and storage) could be automated at an estimated cost of $16, 600, 000 for hardware and $11, 600, 000 for software. The existing files, including the National Union Catalog, could be converted at a cost of approximately $3, 331, 000. Thereafter, the expected annual costs would be $4, 168, 000 for operating salaries, and $317, 600 in recurring expenses for materials, maintenance, space rental, etc.

2. ORGANIZATION X (for abstracting services)

In proposing to Congress "A National Plan for Science Abstracting and Indexing Services,"[21] Heller Associates suggest the possibility of coordinating the major abstracting services (nearly 300) now covering science and engineering technology. Of these services, 18 are profession-oriented (sponsored by large professional societies in most cases) and tend to be stable because of continuous interest in the subjects covered. The bulk of the services (270) are project-oriented, variously sponsored, and may operate only while a new field is being developed.

The proposed joint venture, Organization X, could strengthen the existing services and add new project-oriented abstracting services. It is estimated that such an organization could save millions of dollars a year by eliminating duplicate abstracting, and at the same time improve the coverage of specific projects. Organization X would have access to the materials of the profession-oriented services, would route relevant abstracts to the appropriate project-oriented services, and would provide new project-oriented service as needed when no other sponsor is available.

B. NEW SYSTEMS

Of the five new information storage and retrieval systems reviewed, three could be implemented speedily if approved; two require extensive long-range planning. In the first group are the National Science Information Center, the National Library of Science System and Network (Department of Health, Education and Welfare) and the Interlinking Information Retrieval Networks proposed by Jonker. The more comprehensive systems, still being debated but already influential, are Project PASSIM (for "President's Advisory Staff on Scientific Information Management") and the USSR Data Collection and Economy Planning Network.

1. NATIONAL SCIENCE INFORMATION CENTER (Chicago)

The National Science Foundation, a government agency, has an established Office of Science Information Service which could be the basis of an expanded and automated service. This could take the form proposed by Congressman R. C. Pucinski, who originally suggested a National Research Data Processing and Information Retrieval Center equipped to receive, store, convert and transmit data--by facsimile, teletype, perhaps closed circuit television--to every university and industry. In effect, the world's knowledge could all be made available through such an outlet.[21,22] The Center would not necessarily store all the documents and detailed knowledge, but would keep such complete records of sources and storage locations that other searching would be unnecessary and retrieval would be greatly simplified. The center would coordinate document retrieval and information retrieval, routing a question about thermophysical properties, for example, to the Thermophysical Properties Center at Purdue.

After extensive hearings on the subject, the Pucinski proposal evolved in the direction of a national system of centers instead of a single center. This seems to be the consensus of many knowledgeable groups. The recent National Colloquium on Information Retrieval,[23] for example, tended to discourage the idea of a single large national center, foreseeing instead a national information system with scientists in each field developing their own segment of the system.

2. THE NATIONAL LIBRARY OF SCIENCE SYSTEM AND NETWORK

Another suggestion (the Warren proposal[24, 25]) would expand the MEDLARS system into a national network including not only the medical sciences (which it now handles automatically) but all the major sciences. Libraries, federal and other, and government contractors would list their own and assigned holdings in the scientific literature. Once the contributions were in proper form (a computer-controlled pool of tapes and microforms), the total collection would be replicated for regional centers (seven or more may be needed). This sub-net would then distribute information to university libraries and other users.

To cover added subjects throughly and quickly, such a system would probably start with scientific journals because they are a specific, compact and discrete segment of open literature, commonly accessible and widely consulted. When these regular publications are incorporated in the system and the art becomes more sophisticated, abstracts and facsimile service could be handled efficiently as well. Books, manuscripts and other types of communications could be dealt with separately at a later date.

The advanced experience of MEDLARS could certainly give such a system a sound foundation. For this reason, the Warren proposal calls for physical expansion of MEDLARS, with the U. S. Public Health Service still responsible for system housekeeping.

For this and other proposed systems, and for advanced systems in operation, one serious cause for delay is the shortage of trained personnel to carry the plans into action. An additional specialist group is needed, and appropriate education programs must be developed rapidly in new or existing schools of library science.

3. INTERLINKING INFORMATION RETRIEVAL NETWORKS (Jonker Business Machines)

In a report prepared for the Air Office of Scientific Research, Frederick Jonker[26] questions both the feasibility and desirability of a massive centralized information service. The alternative he proposes emphasizes decentralization, and delegates many network functions to users of the network. This approach tends to keep the cost per search low enough to encourage widespread use, which leads to steady feedback and continuing improvement of the system. The model information retrieval network which he outlines (in its technical, organizational, and financial aspects) could be made operational at the present time, according to Jonker. He estimates that with the network techniques described, 100,000 individual subscribers could be allowed an almost unlimited number of searches of a total of 500,000 documents (acquired at the rate of 100,000 a year, purged after five years) at a basic cost per individual of $15 to $50 per year.

The report assumes the continuing need for separate discipline-oriented and mission-oriented (project-oriented) networks, with the user probably subscribing to one service of each kind. In the normal course of events, the individual users would be employees of member organizations which would join the network either by agreeing to contribute eligible documents, or by taking out blocs of subscriptions, or by both actions (depending on network policy).

The networks could correspond to the structure of existing professional societies in some cases, and would be virtually independent of each other. They would, however, be interlinked by a National Network Coordination Center, which would (1) set common standards to help keep the activities and products of the various networks compatible; (2) serve as a repository for computer tape records of all documents in all networks; and (3) most important, serve as a clearinghouse and transmission point for documents which should be in more than one network. That is, the Center would check the interests of each network in turn against the current computer tapes for all the other networks, passing on to network A (for example) accession listings, bibliographic data, and microfilm copies of relevant documents reported by networks B, C, D, etc.

As projected, each network would supply the user with a regularly updated index for convenient reference, and would automatically compare all new acquisitions with his known interests to call his attention to likely new items. Local stations (Use Units) would serve about twenty individuals each, and would have microfilm readers or reader-printers, and microfilm files of current titles (or titles with abstracts, or document copies, depending on the extent of local interest). These local units would be backed up by regional Information Centers (perhaps one to every five Use Units) which would handle subscription arrangements, gather new documents for the network collection, supply hard copies when needed, arrange for searches of other network collections, and otherwise assist individual users. Completing the networks would be Support and Control Centers, and policy-making Communication Centers.

Government support seems essential at first for such a plan; Jonker advises that it take the form of a revolving grant-and-loan fund. The government might offer special incentives (more grant, less loan) to the first full-fledged network organized, and government-subsidized private firms could run the networks until they become self-supporting, i.e., profitable. Jonker expects subscription costs to put such networks on a paying basis in a relatively short time. He suggests that the first pilot networks be in university communities where they could be working laboratories for improved methods in information retrieval technology.

Indexing methods would probably differ for the two kinds of networks: Within a discipline, shallow indexing based on precombined term pairs; within a mission or project, deep indexing with many terms to identify inter-discipline relationships in the information. These differences in indexing would lead to differences in output materials in the interests of economy. The discipline-centered network would use conventional computer services and microfilm, microcards, etc. The mission-oriented networks would add punched cards or, possibly, "superimposable" cards which can be economically photographed on 35 mm film. A special reader makes these miniaturized cards readable, and the cost may be justified where the number of index terms is very large and references come from many disciplines. Such a technique (the preparation and use of the superimposable cards) is presented as significant in making this system economical and workable.

4. U.S.S.R. DATA COLLECTION AND ECONOMIC PLANNING NETWORK

Planners in the Soviet Union and in other eastern European countries have one special advantage in developing software to simulate the economy: Prediction and control are inherent in a systematically organized economy as they are not in a free, unregulated economy. This difference may be noted in the recently published studies of interindustry relations in Russia. Interpreting the results of these studies, Levine[27] points out the effect of certain basic differences between Western and socialist economic theory. Traditional Western theory concentrates on final products and primary input, but considers the intermediate events to be decided automatically by the price and profit systems. A socialist system, on the other hand, lacks the autonomous feedback loops of the price and profit system, and the planners must substitute another kind of decision process.

When Soviet planners relied on the material balances method (as they did, in the main, until about 1956), they often could not complete the plan for a given year until after that year had started. More recently, applying mathematical techniques developed in the United States, they have found input-output tables useful for checking the consistency of plans prepared by their older method. Such methods, implemented by computers, offer the prospect of making, checking, and revising economic plans with new ease and speed. Methods now being devised will feed into the system some needed information missing from earlier systems--information reflecting the needs of consumers and the profitability of specific economic enterprises.

As reported by Feigenbaum,[28, 29] the proposed Soviet national network for economic control aims eventually to simulate the total economy of that country. The ultimate rate of the system would be 10^{16} computations per year, which Feigenbaum estimates would require the equivalent of one thousand CDC 6600 computers. With planning, control and calculation thus automated, the USSR foresees the possibility of doubling the rate of development of their national economy. As described by Glushkov,[30] the system will include dozens of major computing centers of superhigh productivity (10^6 arithmetic operations per second each). Feeding and supplementing these computing centers will be tens of thousands of centers with relatively low productivity (30,000 arithmetic operations per second). Such a hierarchy of computers in a national network has been anticipated by a number of economic analysts (compare, for example, the "decomposition" principle suggested by Dantzig[31]).

Although Russian computer technology is not now sufficiently advanced to take full advantage of the new forms of analysis, Russian planners are apparently applying the methodology to good effect. They have, for example, developed a folding matrix modification of the Leontief matrix and the result is an interesting display of secondary features which illustrates the wide usefulness of the Leontief approach. (This approach, and the Russian variations, are explained more fully in the Appendix.)

5. PASSIM (The President's Advisory Staff on Scientific Information)

This project, proposed in a Senate Joint Resolution,[32] aims to develop and organize data and decision-aiding systems for the benefit of the country as a whole, on a basis described as interdisciplinary and interagency, intergovernmental, interindustry, and inter-regional. "Its purpose," Humphrey said, "is to make a beginning by bringing together the information management and decision-aiding techniques developed by the top minds of our country, just as two decades ago we brought together our top scientists to harness the atom."

The proposal has far-reaching technical implications. An "interdisciplinary" approach implies a concentrated attack on the problem of scientific information retrieval; and "intergovernmental" foresees the time when groups of governments (possibly the United Nations) may share scientific and economic information of common concern through decision-aiding systems linked together. Such prospects call for a global communication network designed to integrate independent but compatible systems. Steps in that direction are being taken by European countries in the Common Market. Statistically compatible tables have already been published for five countries, and complete coverage is planned.

A high-priority goal of PASSIM is to understand our national economy--what makes it work, and what can make it work better.

Humphrey asserted that the project would be valuable if it solved only one problem: how to sustain economic growth at a rate commensurate with the needs of our society. Referring again to Fig. 1 (whose unshaded area is the principal concern of PASSIM), we can see how relevant data comes from private business and from government agencies, and how science supplies the decision theory and other kinds of help in the necessary analysis, projection, and control measures. It is generally recognized that an up-to-date Leontief matrix would be the best way of organizing such data. Work has begun on an input-output table based on the data from the census of manufactures for 1963, and the hope is that the table can be developed to account for at least 500 sectors of the economy.[33] One problem arises from the nature of the free enterprise system, where individual firms have the right to withhold proprietary data, and we must find ways to collect the summary data needed without jeopardizing this freedom.

The Humphrey resolution also noted the continuing value of the economic simulation programs of the Office of Emergency Planning, and recommended the expansion (since begun) of its Program Analysis for Research Management. In related developments, the Office of Economic Opportunity is planning simulation studies of the economy, and the Department of Commerce released (in November 1964) the 81-sector input-output table for the year 1958.[34] As interpreted by Leontief,[33] that table is of immediate practical value,* and its publication will encourage the necessary interest in the expanded and more current versions which will follow. As Leontief urges in another article,[35] government-business cooperation can keep the national input-output tables up to date, providing "real-time" marketing data for breakthrough in sales forecasting. Along the same lines, the Corporation for Economic and Industrial Research (C.E.I.R.) established a data base and library of programs for economic simulation and marketing studies. Expanding their present

* Copies of this table in the form presented in reference 33 are commercially available as wall charts from the W. H. Freeman Company, San Francisco. (See Scientific American, October 1965, page 101.

limited services of that kind, C.E.I.R. expects to have comprehensive Leontief matrices (1000 x 1000) available for more detailed marketing studies within a few years.[36]

Inherent in the information management problem addressed by PASSIM is the challenge of recentralization. In business and government, information is needed wherever there is authority to make decisions, and management structure is closely related to the pattern of information needs. One discussion of this problem[37] examines the common information problems of the five hundred largest corporations in the United States. The author (Robinson, of C.E.I.R.) suggests that changes are needed, but that the best answer is neither the "democratization" of industry nor imposed government control (where industry is the captive of political bureaucracy), but rather "a voluntary centralization (via the electronic information net) of industrial capability; the growing of a new kind of organization--a new creature, a new animal giving itself a new skeleton, new muscles, a new nervous system."

V. SUMMARY AND CONCLUSIONS

Information management is a critical problem in four overlapping areas: government, business, the military, and science. Our direct interest is in the latter, particularly in the unclassified portion of scientific information as it affects and is affected by information from the other three major segments of the national economy. Surveying first the current services, we find more than 400 information centers dividing their interest among more than 20 areas of science. The document collections vary from under 10,000 to nearly one million, in some of the more advanced centers. Abstracting services, another intermediary between scientific author and reader, are mainly project-oriented. Of 288 major abstracting services, only 18 are definitely oriented toward a single discipline or profession. This being the case, efficient interdisciplinary communication is essential--aerospace projects, for example, depend on research in medicine, metallurgy, astronomy, and on the integration of information from these diverse sources. The profession-oriented and project-oriented agencies, recognizing their complementary functions, could work to eliminate duplicate effort and to improve the interchange of information among themselves.

Present abstracting services in the United States vary from the conventional to the automated and specialized. Engineering Index, Chemical Abstracts, and Index Medicus (National Library of Medicine) are well-established and profession-oriented; the Nuclear Data Project and the Thermophysical Properties Research Center are more recently-formed project-oriented services. A new government service called Clearinghouse is included as an example of a central communication center for identifying and locating technical reports.

Services outside the United States deserve attention in this study for different reasons. Science Abstracts (England) is of interest to us because it continues to be the primary source of abstracts for American physicists. In the case of VINITI, whose services are available only within the USSR, our interest is in the way traditional library and abstracting techniques have been organized on a massive scale to keep Russian scientists informed about scientific developments throughout the world. The trend in this country is not toward such a centralized system as VINITI, but toward a national network of compatible but independent information storage and retrieval centers for different fields of science.

Advanced systems (i.e., those relying on considerable automation in storing and retrieving information) now operate in several specific fields; they are enlarging the scope of established services; and they offer instructive examples for new systems being considered. The Defense Documentation Center, for example, pioneered in computerized report-processing and bibliographic search on a large scale. The MEDLARS system, most notable for its achievements in the computerized preparation of the monthly Index Medicus, also prepares "demand" bibliographies on a limited basis (but must be improved before it can offer general bibliographic services). The MEDLARS study suggests one important technological target: equipment to handle both digital and video records, so that the system could reproduce complete text including illustrations. Needed more immediately by MEDLARS is a compact high-speed large-capacity memory for more efficient bibliographic searching. Science Citation Index, which now provides virtually complete bibliographic service for the medical sciences, is extending its service to other areas of science, and its method assures unusually thorough coverage.

Pilot systems, operating on a relatively small scale, provide a realistic basis for the comparison of alternate methods and equipment. In SASIDS (NASA), for example, the emphasis is on the quality of the information retrieved. The problem here is how to keep the system aware of the changing interests of the user, how to avoid sending him irrelevant information, how to improve the system on the basis of feedback from the user. To this end, a SASIDS matrix, developed and updated by computer, correlates the interest profiles of the subscribing scientists with the declared interests of authors of incoming documents. This correlation determines the dissemination pattern. Another pilot system, SMART (Harvard) tests the effectiveness of different methodologies in information retrieval, with some interesting if tentative conclusions. The Library of the 21st Century (Bolt, Beranek and Newman) demonstrates on a small scale how both documents and information might be dispensed in the library of the future.

In the proposal stage are a number of systems, some new, some revisions of existing systems. In the second category, one major proposal concerns the Library of Congress, which could be automated by converting the card catalog to machine-readable form. After an initial cost of about 31 million dollars, the annual operating cost is estimated at about 4 1/2 million which compares favorably with the cost of present manual operations (allowing for anticipated increases). An important technological target is presented here, since complete and efficient automation of that catalog would require an economical, compact, trillion-bit memory. Another proposal would realign existing services by establishing Organization X--an interchange for the 18 profession-oriented and 270 project-oriented abstracting services.

Among the proposed new systems reviewed, the first (Congressman Pucinski's HR 1946) originally recommended a National Information Center, but in the hearings generated by the proposal, the concept changed to a national network of communicating centers, each covering one branch of science. Another proposal (by Warren) would establish such a system based on the National Library of Science system and network of the kind that evolved, in concept, from the Pucinski proposal. The Warren proposal would start with the medical branches of the national network by expanding MEDLARS to provide duplicate magnetic tapes and microfilm files of documents for regional branch libraries.

A third proposal (by Jonker) also assumes the desirability of a network of information systems, one for each branch of science, but recommends a specific financial structure and specific hardware. The hardware proposals include feasibility estimates for specific microfilm-based technology, but may be subject to argument in view of the problems of standardization.

The U.S.S.R. proposal for a national data collection and economic planning network is of interest even though it is only indirectly related to scientific information retrieval. Concentrating on inter-industry relations as reported by a hierarchical computer-communication system, the Russian network involves problems common to any inter-disciplinary system and any national network of information centers. As planned, the U.S.S.R. system will place a small computer in virtually every factory, city office, and regional office to continually summarize and report data on production, consumption and other changes pertinent to the economy. These local installations will feed information to several dozen large computers whose output will be the basis for decisions in the central planning offices in Moscow. Software techniques now being applied (suggested by Leontief's work at Harvard, and by linear programming methods developed at Rand Corporation) will doubtless hasten the time when planning officials will have a continually updated view of how the economy is functioning. The Russians aim to double their rate of economic growth by means of the projected computer-communication system.

Coming, finally, to the most comprehensive and most official proposal for a new system in the United States, we find PASSIM (President's Advisory Staff on Scientific Information Management) tackling both the problem of scientific information storage and retrieval, and the problem of monitoring the state of the economy. When PASSIM was proposed by Humphrey in the Senate, the emphasis was on establishing and strengthening information ties between all kinds of interdependent units--political, industrial, academic. Taking Humphrey's key terms as starting points for a search of recent literature, we found (under such terms as "interdisciplinary...intersector...intergovernmental") some convincing corroboration of the needs he stated in the proposal. We noted the publication in November 1964 of the 1958 input-output tables for the United States, and compared this preparation time with that for similar publications in other countries--in France, for example, which maintains a schedule of current estimates.

We found Leontief proposing a way to use input-output tables to generate marketing data for a breakthrough for business management. We found some pertinent questions being raised about "recentralization": how can the United States recentralize to benefit from modern computer technology without doing violence to the free enterprise system, i.e., without copying the central control of the Russian system. A marketing service, announced in March of this year, offers computer analysis of information of the national input-output tables. In these and other ways, then, portions of Project PASSIM are taking shape on the initiative of individual organizations. In line with these developments, the predominant recommendation at a recent conference on the subject[23] was for a loosely-knit national information system with scientists in each field developing a segment of the system.

Long-range plans are being projected by another group, the Committee on Scientific and Technical Information (COSATI), which represents the interests of all federal agencies.[38] Working under an assignment from the Federal Council for Science and Technology, COSATI is subdivided into eight panels, each studying one aspect of the problem: Information Generation, Operational Techniques and Systems, Information Users, Information Processing Technology, International Activities, Budgets and Statistics, External Relationships, and Education and Training. A COSATI report scheduled for fall publication is expected to lead to comprehensive recommendations for government action and to related responses from non-government organizations.

In the course of this study, we noted some interesting analogies between the management of scientific information and the management of economic and business information. For example, items listed in the Science Citation Index can be viewed as row entries in a matrix of input-output relationships between scientists, as the input-output relationships between industries are the row entries in a Leontief matrix. In both cases, the transactions represented are already a year or two past when they are entered. Such undesirable time lags, in these and other information systems, are a direct challenge to computer and communications technology, especially when real-time data collection is being planned for an entire economy as large as that of the U.S.S.R.

On a much smaller scale, SASIDS also relies on a matrix form of analysis, attempting to estimate the probability that Scientist A would cite the work of Scientists B and C if he knew of their work. Given unlimited output of scientific information, and limited interest on the part of each scientist, the practical system needs to disseminate information much more selectively than it stores information. (Too many wrong numbers, or misdirected letters, or misleading headlines can make any "service" a nuisance instead of a help.) Confidence and participation in the system depend on a high level of usefulness in the information distributed, but that level can be reached and maintained only by reliable feedback from system users. The government might find it worthwhile to subsidize the cost of terminals and communication channels in such a system for scientists who would help in the experimental stages--entering abstracts, replying to questions, etc. Even so, this is likely to remain a critical problem, and solutions for large systems may well be guided by the results in small pilot systems such as SASIDS.

Similarly, each of the other systems approaches the problem of scientific information storage and retrieval from a slightly different angle; taken together they offer some instructive lessons about how this information is now transmitted, and how it might be made more accessible and more timely.

ACKNOWLEDGMENTS

The course of this study was guided by the initial suggestions of Dr. E. Hopner, and by valuable discussions with Dr. G. K. Machol and Dr. C. J. Hoppel. The author is also indebted to the late Dr. E. A. Quade, and to Mr. T. D. Phillips and Dr. A. Warheit for constructive reviews of this report.

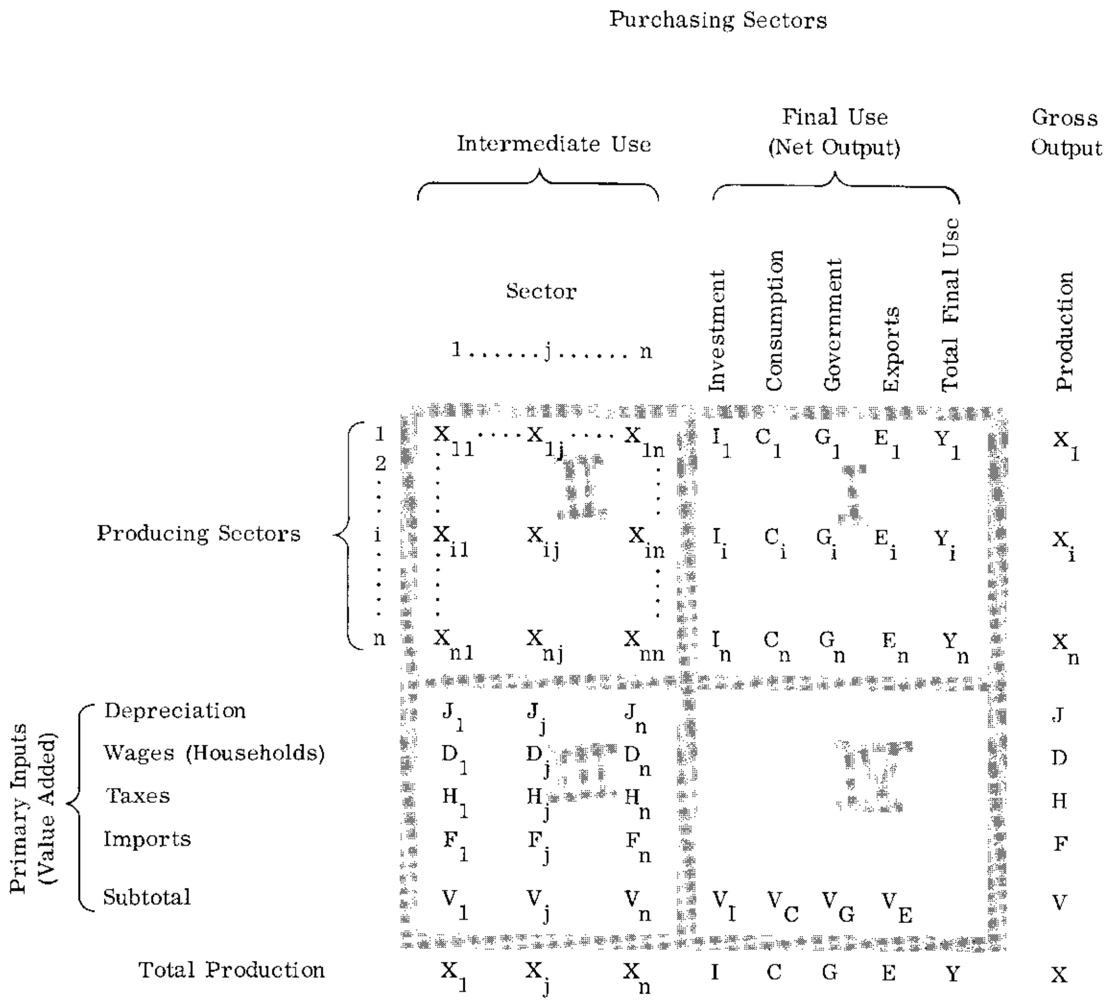

Fig. 3. Adaptation of Leontief Input-Output Matrix[40]

APPENDIX: INPUT-OUTPUT ANALYSIS IN THE UNITED STATES AND IN THE U.S.S.R.

The chief "information users" in a country such as the United States--the scientific community, private industry, government agencies and the military--are interdependent in the sense that decisions in any of these groups call for information about the activities and plans of the others. For this reason, our interest in scientific information storage and retrieval leads to an interest in methods of summarizing data from all these groups for the benefit of each. Ideally, information of common importance should be accessible, up-to-date, and organized in a format which analyzes the data it displays: input-output analysis makes this possible. This method can be applied to the national economy, concentrating on all the factors influencing production, so that the structure of dependence, being understood, can be altered purposefully when that is needed.

From the historical point of view, the input-output method stems from general equilibrium analysis, customarily traced to the French physiocrat Francois Quesnay, whose Tableau Économique was published in 1758. In 1877 the French economist Leon Walras produced an abstract mathematical model based on the interdependencies of the productive sectors of the economy. In the mid-1930's, Wassily Leontief (Harvard University) developed the first applied model of this type, working with empirical data for the economy of the United States.[39] Since then, research in input-output economics has grown rapidly, and statistical analysis of interindustry relations has been undertaken in more than twenty countries to complement the more traditional forms of economic studies. Such interindustry studies can be valuable in guiding both private and public bodies in investment decisions and other matters related to resource allocation.

Prior to 1964, the interindustry study for 1947 was the last officially published for the United States, formal government work of this kind being suspended in 1952. In November 1964, however, the results of a study based on 1958 data were made available by the United States Department of Commerce,[35] and this study took the form of Leontief input-output analysis.

The Leontief analytical system is based on an input-output table which shows how the output of each industry is distributed among other industries as input for further processing, and how the output finally reaches consumers, as goods or services. The entries in an input-output table are normally the dollar values of all industry purchases and sales (as reported in census data). The basic assumptions are that, for any particular industry, dollar values are a quantitative measure of physical purchases per unit of output, and that this pattern remains relatively undistorted by long-run price changes. The soundness of this method also depends on the fact that an industrial society changes rather gradually; i.e., the dollar value matrix of transactions for one year will not vary grossly from the pattern for the year before or the year after (although the gross output to final consumption may vary).

Before an input-output table is prepared, decisions must be made as to its ultimate size and how to combine industries in representative production sectors. (The intended use of the table will guide these decisions.) Taking the firm as the basic business unit, and the industry as the class of firms producing similar goods and services, we proceed to the "sector," a useful concept for analytical purposes. Input-output tables may be comprised of aggregates such as the agricultural sector, the manufacturing sector, and the household sector, for example. The term "sector" may include government operations, foreign trade and capital formation; a "final demand" sector may encompass all groups which purchase final products.

Essentially, the practical problem for the input-output analyst is to reduce the number of elements to be considered from several thousand (the number of individual industries) to a more manageable number. Once the size and sectoral composition of the tables are settled, the matrix of interindustry transactions in dollar values is prepared as an input coefficient table (showing direct industry purchases per unit of output). Converted into an inverse matrix (the Leontief inverse), this shows the direct plus indirect industry purchases necessary for a unit increase in industry output to final purchasing sectors.

To see how this works, let us examine Fig. 3.[40] Element X_{ij} (in quadrant II) indicates the quantity of goods produced by sector i for intermediate use by sector j. To determine the total cost of sector j production, we add the elements in column j down through quadrant II (materials, semi-finished parts, subassemblies purchased from other industries) and through quadrant III (depreciation, wages, taxes, and cost of imported materials). To determine the disposition of products from a given sector, we read across quadrant II (elements representing the sale of sub-assemblies and parts to other industries) and across quadrant I (the distribution of output to final use). The totals of the corresponding columns and rows will agree, but the table total is not equal to the gross national product, since most products are counted twice here (first as raw material, etc. and then as part of a product in the final use quadrant).

As may well be imagined, the fact-gathering task is the biggest part of the job of preparing the input-output table. Cooperation between government and industry--establishing consistent methods of reporting, for example--could obviously reduce the time required, and keep the tables more nearly current. The task of computation, which is also immense, works from this premise: "all possible interconnections...are special instances of the general solution of a single large system of equations in matrix algebra."[33] Computing the 81-sector table of transactions (for 1958, released in 1964) thus required the solution of 81 equations each involving 81 distinct but interdependent variables. These solutions were carried out to five significant figures in three minutes of IBM 7090 time. The 500+ sectors planned for the next table will place no undue burden on the computers, but gathering and organizing that much information will clearly require concerted effort.

Such a detailed view of the economy is in effect a working model whose components can be varied in numerous ways for experimental purposes. As Leontief says, the approach "tends to make the analyst consider many aspects of the economic process, and above all to consider them in a mutually consistent way."[33] The result should be a sounder basis for marketing analysis, resource allocation, and other prediction and control activities of industry and government.

In a socialist economy such as the U.S.S.R., the problem differs in the greater degree of government control, which leads to standardization in gathering information and generally precludes problems of proprietary or confidential information. Another difference is the absence of consumer feedback (although this appears to be changing) and the more limited computer technology available (this will change too, of course). Working within these particular opportunities and restrictions, Russian planners have developed additional applications and interesting displays of

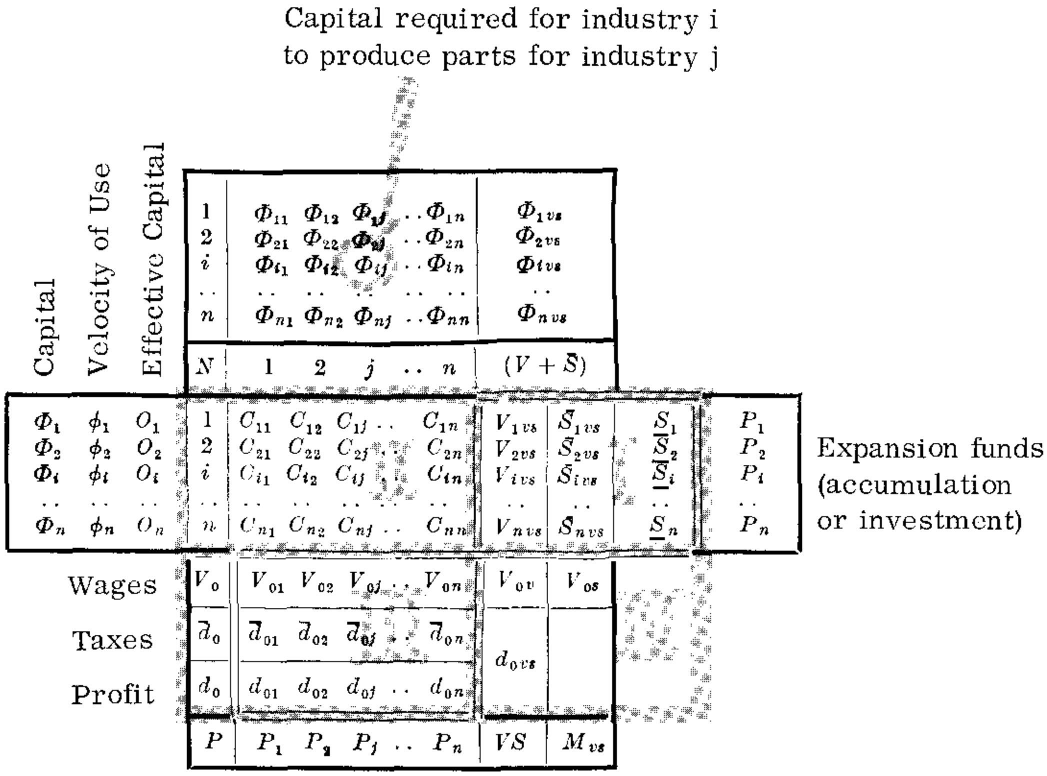

Fig. 4. Leontief Matrix with Added Wings for Capital Requirements*

*Adapted from Oparin[41] and reproduced with permission of Pergamon Press.

secondary features which originated with Leontief at Harvard. That is, they are applying the input-output method modified to suit their purposes, and have worked out a "folding matrix" which offers an overall perspective of the economy in a relatively simple form.

In a sample of the folding matrix (Fig. 4), the heavily-outlined square corresponds to the matrix in Fig. 3, and the central section (C_{ij}) is the same as quadrant II. As before, the costs of industry j are added vertically through quadrants II and III. Reading down, the cost (P_j) is $C_{1j} + C_{2j} \ldots + C_{nj} + V_{0j} + J_{0j} + d_{0j}$. The last three are wages, taxes and profit, respectively. (The subscript zero marks monetary transactions as distinct from deliveries of materials and products.) The total production is accounted for by reading horizontally $C_{i1} + C_{i2} \ldots + C_{in} + V_{ivs} + \overline{S}_{ivs} + \underline{S}_{ivs}$. The last three represent direct public consumption, state expenditures (such as our public works), and accumulation or investment for expansion.

The fourth quadrant, part of the non-productive sector, includes wages for direct public services (V_{0v}) and for administrative services (V_{0s}); and an additional term (d_{0vs}) covering license fees, special taxes, and any profits accrued from public services.

Coming to the added wings of the Russian folding matrix, we find that they prescribe, interestingly enough, a clear and simple way of tabulating capital requirements. In the left wing, the total capital (Φ_i) times the circulation rate (ϕ_i) equals the effective capital available per year (O_i). The effective capital must equal the combined value of components purchased from other sectors, goods delivered for direct consumption, and goods supplied for government construction. The wing added to the top of the matrix expands the column of capital requirements per sector into a full matrix (Φ_{ij}) of intersector capital requirements. Column Φ_{ivs} represents the capital required for the final processing, assembly and delivery of goods to final consumption and government service.

In a numerical example (Fig. 5), we see how the folding matrix provides a perspective of the total economy. Calculated in billions of rubles, the total outputs of the rows across quadrants II and I (80 + 70 + 10 + 40) must equal the total inputs of the columns down through quadrants II and III (80 + 60 + 20 + 40). Similarly,

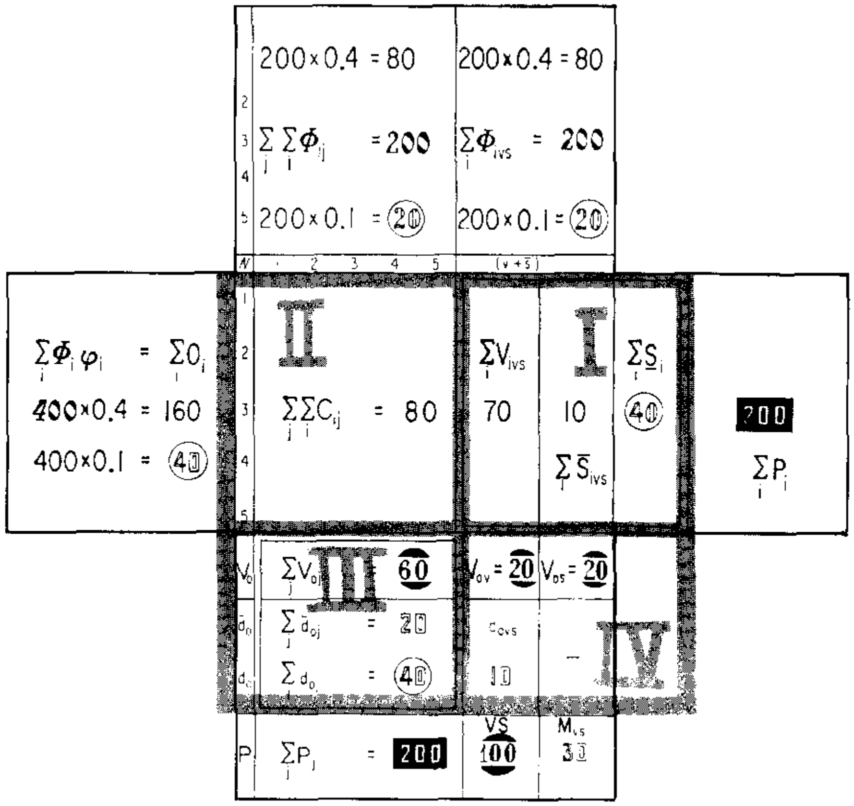

Fig. 5. Numerical Example (in Billions of Rubles) of Folding Matrix Analysis of Economy*

*Adapted from Oparin[41] and reproduced with permission of Pergamon Press.

consumption, added vertically in quadrants I and IV (e.g., 70 + 20 + 10) must equal wages, added horizontally in quadrants III and IV (60 + 20 + 20). Individual lines and corresponding columns will not necessarily balance with national income, but consumption and accumulation added horizontally (70 + 10 + 40) must equal net production, i.e., total production less the intermediate material expenditure (200 - 80).

The matrix allows numerous ways of checking various features of the economy by means of equations suggested in Leontief's basic work or implicit in his approach. Accumulation (or investment) balances added vertically ($\underline{S}$ = 40) must equal profits added horizontally (d_0 = 40). The total capital funds in the left wing (Φ = 400) must equal the sum of the capital in the two upper wings ($\Phi + \Phi_{vs}$ or 200 + 200). The funds at the beginning of a period (e.g., 400) times the velocity of circulation (0.4) must equal the sum of production (80 + 70 + 10) and the (vertical) sum of intermediate production, wages and taxes (80 + 60 + 20). Again, the total capital funds (400) times the desired rate of expansion of the economy (0.1) must equal the accumulation $\underline{S}$ = 40. Each sector can be further analyzed on the basis of speed of circulation and rate of growth. The total production (200) must equal the total capital (400) times the sum of the velocity of circulation and the rate of growth (0.4 + 0.1).

As we study the folding matrix, other cross-checks on the balancing of accounts become apparent, and these modifications developed in the U.S.S.R. for displaying the relationships may suggest useful new ways of applying the matrix method to the somewhat different problems of our own economy.

REFERENCES

1. "Science, Government and Information," (A Report of the President's Science Advisory Committee, Panel on Science Information, A. M. Weinberg, Chairman), U.S. Government Printing Office, Washington, D.C., January 10, 1963.

2. Weinberg, A.M., "Second Thoughts on Scientific Information Retrieval," College and Research Libraries, November 1964, pp. 463-471.

3. "Specialized Science Information Services in the United States (Physical and Biological Sciences)," National Science Foundation No. 61-68, November 1961.

4. Kiell, N., Psychoanalysis, Psychology and Literature, A Bibliography, (University of Wisconsin Press, Madison, 1963), 225 pp., 4460 entries.

5. Ruttenberg, C.L., "Report on Data Archives in the Social Sciences," The American Behavioral Scientist, Vol. VIII, No. 8, April 1965, pp. 34-35.

6. Herling, J.P., "Engineering Abstracting Services," Special Libraries, December 1961, pp. 560-565.

7. Cotrell, N. E., "Information Retrieval: Where is it Heading?" The Engineer, Vol. 6, No. 2, Engineers' Joint Council, New York, 1965.

8. Gove, N.B., "Information Centers in Nuclear Physics," Physics Today, January 1964, pp. 44-47.

9. Cezairliyan, A.O., Lykoudis, P.S., and Touloukian, Y.S., "Analytical and Experimental Study of a Method for Literature Search in Abstracting Journals," TPRC Report 11, Purdue University, December 1960, OTS-171478.

10. "A Brief Description of VINITI," Hearings on House Resolution 1946, Appendix to Vol. 1, 1963, pp. 373-409.

11. Kridler, T., and Simpson, G., "Should Science Information be Centralized or Decentralized?" The American Behavioral Scientist, Vol. VIII, No. 6, February 1965, pp. 29-33.

12. Austin, C.J., "The MEDLARS System," Datamation, December 1964, pp. 28-31.

13. "The MEDLARS Story at the NLM," U.S. Department of Health, Education and Welfare, Public Health Service, 1963, pp. 61-62.

14. Garfield, E., "Science Citation Index--A New Dimension in Indexing," Science, May 8, 1964, Vol. 144, No. 3619, pp. 649-654.

15. Drèze, J.H., "Some Postwar Contributions of French Economists to Theory and Public Policy, with Special Emphasis on Problems of Resource Allocation," American Economic Review, Vol. LIV, No. 4, Part 2, Supplement, June 1964, pp. 1-64.

16. Mann, L.D., and Pillorgé, G.J., "French Regional Planning," Journal of the American Institute of Planners, Special Issue, Vol. XXX, No. 2, Regional Development and Planning, May 1964, pp. 155-160. (Includes short bibliography.)
See also: Hackett, J. and A., Economic Planning in France (Harvard University Press, Cambridge, Massachusetts, 1963).

17. Salton, G., "The Evaluation of Automatic Retrieval Procedures--Selected Test Results Using the SMART System," American Documentation, Vol. 16, No. 3, July 1965, pp. 209-222.

18. Salton, G., "Information Storage and Retrieval," Computation Laboratory of Harvard University, Scientific Report No. ISR-8, to National Science Foundation, December 1964.

19. Licklider, J.C.R., Libraries of the Future (M.I.T. Press, 1965).

20. King, G.W., editor, Automation and the Library of Congress (U.S. Government Printing Office, Washington D.C., 1963).

21. National Information Center: Hearings ... on House Resolution 1946, Appendix to Volume 1, Parts 1, 2, and 3 (U.S. Government Printing Office, Washington, D.C., 1963), pp. 109-153.

22. National Information Center: Hearings ... on House Resolution 1946, Volume 1, Parts 1, 2, and 3 (U.S. Government Printing Office, Washington, D.C., 1963).

23. Proceedings of the Second Annual National Colloquium on Information Retrieval, April 23-24, 1965, University of Pennsylvania. (To be published.)

24. Warren, Dr. S. L., private correspondence, February 10, 1964.

25. Science Newsletter, Vol. 85, p. 371, June 13, 1964.

26. Jonker, F., "A Model Information Retrieval Network for Government, Science and Industry," Information Research Division, Air Office of Scientific Research, AF 49-(638)1209, May 1964 (AD600221).

27. Levine, H.S., "Input-Output Analysis and Soviet Planning," American Economic Review, Vol. 61 (No. 2), May 1962, p. 127.

28. Feigenbaum, E.A., "Russian EDP Used to Boost Government Economy Control," Electronic News, December 14, 1964.

29. Feigenbaum, E.A., seminar, Stanford University, January 6, 1965.

30. Glushkov, V., et al., "Certain Problems of Cybernetics," Izvestiya, Moscow, September 6, 1964, p. 4. (Translation: JPRS 27440, November 19, 1964; Clearinghouse TT: 64-51689).

31. Dantzig, G.B., Linear Programming and Extension (Princeton University Press, 1963). See especially pp. 448-470.

32. Senate Joint Resolution 202, September 10, 1964, Congressional Record-Senate, pp. 21281-5. (See reference in Legislative Bulletin, National Society of Professional Engineers, October 1964.)

33. Leontief, W.W., "The Structure of the U.S. Economy," Scientific American, April 1965, pp. 25-35.

34. Goldman, M.R., et al., "The Interindustry Structure of the United States--A Report on the 1958 Input-Output Study," U.S. Department of Commerce, Office of Business Economics, Survey of Current Business, Vol. 44, No. 11, November 1964, pp. 10-29.

35. Leontief, W.W., "Proposal for Better Business Forecasting," Harvard Business Review, November-December 1964, p. 166 ff.

36. "Input-Output: Computer-Age Top Management Tool for Business and Government," C.E.I.R., Arlington, Virginia.

37. Robinson, H.W., "The Challenge and Promise of Recentralization," C.E.I.R., Washington, D.C.

38. "Long-Range Information Plan to be Developed," Scientific Information Notes (National Science Foundation), June-July 1965, pp. 2-3.

39. Leontief, W.W., Structure of the American Economy (Oxford University Press, New York, 1951).

40. Lofting, E.M., and McGauhey, P.H., "Economic Evolution of Water. Part III: Interindustry Analysis of the California Water Economy," University of California Water Resource Center, No. 67, January 1963.

For a more generally available example of interindustry accounting, see Chenery, H.B., and Clark, P.G., Interindustry Economics (John Wiley, New York, 1959), p. 16.

41. Oparin, E.I., Multi-Sector Economic Accounts (Pergamon Press, New York, 1963) pp. 19-20.