INFORMATION RETRIEVAL

A Critical View

Edited by GEORGE SCHECTER

Based on Third Annual Colloquium
on Information Retrieval
May 12-13, 1966, Philadelphia, Pennsylvania

Sponsored by

Special Interest Group on Information Retrieval,
Association for Computing Machinery

Delaware Valley Chapter, Association for
Computing Machinery

Institute of Electrical and Electronic Engineers
Computer Group, Philadelphia Section

United States Army Frankford Arsenal

Moore School of Electrical Engineering,
University of Pennsylvania

1967

THOMPSON BOOK COMPANY
Washington, D. C.

●

ACADEMIC PRESS
London

THOMPSON BOOK COMPANY
383 National Press Building
Washington, D. C. 20004

Exclusive sales rights for this book throughout the entire world (with the exception of North and South America):

ACADEMIC PRESS
Berkeley Square House
Berkeley Square, London, W.1

Library of Congress catalog number: 67-29616

PRINTED IN THE UNITED STATES OF AMERICA

OFFICE OF THE VICE PRESIDENT
WASHINGTON, D. C.

April 29, 1966

Dear Mr. Schecter:

America does need to take a critical view of where we stand in information retrieval. Fortunately, as you know, a great deal of progress has been made since the Space Age dawned in October, 1957. In government, in the nation's universities, in private industries, in the scientific and technical press and in many other realms, impressive advances are occurring at this very time.

Yet, today more than ever before, the deluge of information in every form — printed, written, visual, audio and oral, words, numerals and illustrations, manual and machine — in all sorts of permutations and combinations — represents one of the greatest challenges and opportunities to this dynamic era.

I would appreciate, therefore, if you would convey my greetings to the Third Annual National Colloquium on Information Retrieval. The distinguished experts who will address your sessions will, I know, provide real stimulus to further advances. Your audience will respond not only on the scene, but in follow-through as they return to their respective organizations and locales.

Above all, we need today that voluntary teamwork which will build information networks, rather than merely proliferate an endless number of separate entities. A new type of excellence is needed — excellence within and between the respective "universes" of information, rather than merely excellence within a particular system or subsystem.

In every aspect of human affairs — science, technology, economics, culture, politics — information is exploding, breaking through traditional lines of demarcation, and requiring use, re-use and re-routing.

President Johnson and his Administration will continue to be a catalyst for information progress and cooperation. But in the final analysis, what is achieved in the private sector, both by non-profit and profit-making organizations, will provide the decisive breakthroughs.

My best wishes for fruitful sessions.

Sincerely,
Hubert H. Humphrey

THIRD ANNUAL NATIONAL COLLOQUIUM ON INFORMATION RETRIEVAL

COMMITTEE

GEORGE SCHECTER, Chairman
United States Army Frankford Arsenal
ALBERT TONIK, Vice Chairman
UNIVAC Division of Sperry Rand Corporation
ASHLEY SPEAKMAN, Secretary
E. I. du Pont de Nemours & Company, Inc.
RICHARD S. FRARY, Treasurer
Ultronic Systems Corporation
MORRIS RUBINOFF, Program Co-Chairman
Moore School of Electrical Engineering, University of Pennsylvania
SYLVAN H. EISMAN, Program Co-Chairman
United States Army Frankford Arsenal
GERALD L. BRODSKY
Auerbach Corporation
REGINA HILDRETH
Radio Corporation of America
DAVID LEFKOVITZ
Moore School of Electrical Engineering, University of Pennsylvania
THOMAS C. LOWE
Moore School of Electrical Engineering, University of Pennsylvania
HENRY SPARKS
Moore School of Electrical Engineering, University of Pennsylvania

PREFACE

The growing importance of information accessibility has been expressed in many ways by many authorities. It has been agreed generally that the accelerating growth of the world's information and of its generators and users are reasons enough to warrant improved retrieval techniques. Coupled with this is the increasing recognition that soundly based decision-making processes in all human affairs, public and private, are dependent upon analyses of information from numerous sources and diverse fields. The consequences of inadequate or incorrect information in defining decision options can be serious.

The critical resources one deals with in these processes are INFORMATION and WISDOM. Information is growing at an unprecedented rate; both the number of its possessors and the total quantity of information are multiplying geometrically. Wisdom, the ability to judge soundly and deal sagaciously with facts, is a much rarer resource — not growing noticeably — hopefully not shrinking. Therefore, let the information scientist bend his efforts to the formidable task of devising means of making the information available in the expectation that it will reach him who will acquire knowledge — and in the hope that it will reach him who will use it wisely.

These proceedings of the Third Annual National Colloquium on Information Retrieval are statements of today's pathfinders, exploring new ground at the frontiers of this vital field. To be a pioneer in intellectual endeavors is commendable enough — but to express new ideas from a public platform, to commit them to print, and to expose them voluntarily to a sophisticated group of one's peers is an act of courage and conviction. This is the stuff our speakers and authors are made of — and they deserve warm thanks for making the Colloquium and this book possible.

No pretense is made here that the field is covered systematically or comprehensively; rather we have selected for examination a few outstanding examples of operating and projected retrieval techniques and system types. The result is a broad sampling of representative approaches to design philosophy, use technique, system implementation, and system performance evaluation.

The Colloquium was made possible by the voluntary efforts of Mr. Albert Tonik, Mr. Ashley Speakman, Mr. Richard S. Frary, Dr. Morris Rubinoff, Mr. Sylvan H. Eisman, Mr. Gerald L. Brodsky, Mrs. Regina Hildreth, Dr. David Lefkowitz, Dr. Thomas C. Lowe, and Dr. Nathaniel Kornfield.

GEORGE SCHECTER

Feasterville, Pa.
21 December 1966

MOVING CONGRESS INTO THE AGE OF THE COMPUTER

Congressman William S. Moorhead (D., Pa.)

Today men stand in awe — some in fear — of the great computing machines.

We know our own frailties. The great machines appear to have none.

People wonder about the relationship of men and machines. Who is master? Who is servant?

Some lines of the great Poet Laureate John Masefield may help us decide. He said:

> Man consists of body, mind and imagination.
> His body is faulty, his mind untrustworthy,
> but his imagination has made him remarkable.

The story of the great mathematician, Karl Friedrich Gauss, illustrates this precept. In the year 1809 his imagination led him to develop the formula for computing the orbits of the planets in the solar system.

Subsequently he spent 20 years computing the orbits of the various planets. His greatness depends not upon his 20 years of computations, but upon his one remarkable formula. Today Gauss' 20 years of computations could be accomplished in less than a week on a modern computer. Imagine to what greater heights the imagination of this great mathematician might have soared had he had a computer to do his unimaginative work of calculation.

So I stand here now, not so much in awe of computers as in awe of you, the information scientists, who have the *imagination* to *devise* the *in*put and who have the *imagination* to *use* the *out*put of these great machines.

I suppose that I am here today because of my sponsorship in this Congress of the bill first introduced several years ago by the then Senator Hubert H. Humphrey — the bill which would establish in the Executive Office of the President the *President's Advisory Staff on Scientific Information Management* (or PASSIM). The PASSIM bill is now pending before the Government Operations Committee, of which I am a member.

I suppose that I am here also because of the Automatic Data Processing (ADP) bill which was enacted by the Congress last year. This is the bill which seeks the most economic and efficient use of ADP equipment by federal departments and agencies. It was drafted by the Government Activities Subcommittee of the House Committee on Government Operations.

My study of both the PASSIM and ADP bills has taught me something about the quantity and cost of ADP equipment in the Executive Branch of our federal government and the importance of having an advisory group at the presidential level.

The growth of ADP facilities within the federal community during the past decade has been tremendously significant. In 1956 there were only 90 computers. Today there are more than 2000 computers in use in some 35 government agencies.

As federal agencies have gained more experience with ADP systems, there has come a better understanding of how to plan their integration into existing operations. Lessons learned in one quarter often are adaptable elsewhere.

Let me point to a few examples in the federal government of orderly and coordinated programs.

Probably the most highly publicized use of computers in the federal community is that in the Department of Defense. Because technological requirements related to national security and military preparedness accelerated the recognition of priority applications for ADP support, the Defense Department was obliged to conduct research and development in ADP-oriented techniques, procedures and systems. Early emphasis was put on the command-and-control area for such groups as the Strategic Air Command. There was also an early substantial effort to automate logistics and intelligence data in support of various service echelon commanders. Government-contractor teams of hundreds of engineers, systems analysts, programmers, and military personnel worked to provide rapid response systems essential to our defense posture. Large and small computers, fixed and mobile, were developed. Input/output devices for remote questioning of data banks by paper tape, punched card, magnetic tape or other media were put into operation. It also became feasible to have point-to-point transmission within a far-flung network — often from computer to computer. And so the manipulation and use of alphabetic information, as well as numeric data, became possible as the result of intensive work on many Defense Department projects.

At the present time, ADP functions are performed for the Department in such major areas as accounting and finances, personnel and manpower, materiel and equipment, maintenance and repair, research and development, operations and intelligence.

As the field commands and headquarters have become aware of the usefulness of ADP, existing systems have had to take on additional requirements for support. Thus, additional capability has been justified on the basis of demand for services. This, of course, has been the story of the growth of ADP equipment throughout the country.

The Defense Supply Agency (DSA) is heavily involved in areas dependent upon knowledge retrieval. DSA has such high-volume data handling operations as the processing of requisitions, maintaining stock records and procurement status, managing materiel, and billing and collecting. The Defense Logistics Services Center data processing records are the world's largest collection of strategic elements of supply intelligence.

The Defense Documentation Center has the task of keeping at hand all of the research results obtained by or for the military services. Each incoming technical or scientific report is indexed and the descriptors are placed on magnetic tape for future computer manipulation. On request — either by a government user or qualified contractor — the computer prints out cards describing and identifying the relevant reports.

One of the first federal agencies to turn to ADP was the Bureau of the Census. With several types of computers, Census collects and processes statistical information on population, housing, agriculture, business, construction, foreign trade, government, industry and transportation.

With its computer facilities centralized in the Washington area, the Bureau of Labor Statistics of the Department of Labor collects, analyzes, and publishes national, regional, and local data on retail and wholesale prices, consumer spending, wages and salaries, employment and unemployment, productivity and technological development, and average hours of work and earnings.

The Social Security Administration began using computers back in 1955. Today, Social Security uses ADP — at a large installation in Baltimore — to keep track of basic payments to all our citizens who qualify, to receive new claims for payment in addition to those from persons already qualified, and to record changes in status of the claimants.

I'm sure you are aware too that the Internal Revenue Service (IRS) has developed a large-scale ADP capability in processing income tax returns.

Of the 104 million returns filed in 1966, more than half will be handled by ADP. The entire processing system will be computerized by next year.

I'm proud to note, by the way, that the IRS cooperates with the University of Pittsburgh in my own Congressional district in the ADP handling of IRS-oriented legal literature. This cooperative project with the University of Pittsburgh has resulted in placing the Internal Revenue Code and Regulations on magnetic tape. But the University of Pittsburgh center also provides automated law searching of Pennsylvania Supreme Court cases, Pennsylvania Superior Court cases, Pittsburgh city ordinances, Pennsylvania statutes, New Jersey statutes, New York statutes, and the U.S. code. And the Center is continuing to expand its library so that as time goes on the computer files at the University of Pittsburgh will become increasingly more useful to the legislative, judicial, and executive functions of government as well as to the business community and the private practitioner. This Center was described to me recently by Dr. Edward Wenk, Chief of the Science Policy Research Division of the Library of Congress, as an "outstanding capability." He said there is "nothing like it in any other place."

While I am discussing the ADP capabilities in my own congressional district let me also point to the progress made in this field by the Carnegie Institute of Technology.

The Institute's outstanding program has already shown that, in business, computers can help make many managerial decisions. Computer programs exist, for example, to assist middle and upper level executives to develop and evaluate equipment designs, to locate warehouses, to select advertising media, to estimate the reception for new products from consumer survey data, to schedule production and inventories, to select stocks and bonds for trust portfolios, to determine bids and prices, and to monitor and control the operation of complex and continuous production systems.

I think that the techniques which have been developed at Carnegie Tech for solving problems of business can be used to help solve the problems of government.

A brief study of the outstanding uses of computers in the executive branch of our government leads me to two conclusions:

1. To oversee, to correlate, and to utilize at the very highest level of government there should be an advisory group reporting directly to the President. This is the purpose of my bill to establish the President's Advisory Staff on Scientific Information Management — PASSIM.
2. The time has come for Congress to enter the computer age.

On this second point I ask your help.

The future of representative democracy, the future of our constitutional government with its delicate system of checks and balances requires that, in the computer age, the legislative branch of government make full use of computer capability.

Today except for one small unit which the Library of Congress uses to handle its payroll, the Congress of the United States does not own one penny's worth of ADP equipment.

When I tell you this I am expressing my concern for the future of representative government in the United States.

Secretary of Defense McNamara, over at the Pentagon, can conclude, with the help of computers, that it should be our national policy to phase out our bombers, that it should be our national policy *not* to deploy an antimissile system.

But how can the Congress agree or disagree with him? How can the Congress provide or deny him the necessary funds, when the information on which the legislative decision is made comes out of horse-and-buggy procedures?

I invite you on your next visit to Washington to look in on the musty document room in the Capitol where papers are handled in about the same way they were when George Washington was President. Bills are filed away in rusty old metal boxes in floor to ceiling slots, accessible from an old oaken ladder that slides sideways on rollers. We are also so old-fashioned in our ways that we cannot even make a change in our employees payroll roster after the 10th of the month. Payday follows some 20 or 21 days later at the end of the month.

Within the federal government there is a broad spectrum of requirements for information of the right scope and nature. This is equally true in principle, although with modifications in detail, in the Executive, Legislative, and Judicial branches. The problem is seldom a scarcity of information. The three branches of government need *equal access* to facts, since the interplay between them is affected significantly by the degree to which each has access to, and can use properly, vital information.

What is needed, if I may borrow a phrase from another era, are "separate but equal" facilities!

When I spoke a moment ago of my concern for the future of representative government in the United States, I was referring in particular to the *balance* in the relationship between the Executive Branch and the Congress.

The balance is in jeopardy as the result of demands upon the time and energies of the members of Congress.

I propose that Congress now move into the age of the computer.

I propose that the Legislative *Reference* Service of the Library of Congress — a highly-skilled research organization — be equipped with the tools that will enable Congress to move into the computer age.

As a matter of fact, I think it would be a good idea to change its name to the Legislative *Research* Service, as more descriptive of its role now and in the future.

Specifically, we need on the Hill a Central Read-Out facility that could tap the memory banks of all the other computers in the federal government, not to secure privileged data but to secure *facts* such as economic statistics, demographic profiles, and figures on such things of daily concern to a Congressman's office as funds allotted to his district, and project and contract awards.

I think the Federal Budget should be put on a computer for ready access by the Congress.

There should be read-out devices or closed circuit television screens in the offices of Senators and Representatives and in committee offices linked, of course, to the Legislative Research Service.

We need on the Hill too a bibiliography device that will print out a list of references automatically.

And, most of all, we need the skilled staff in the Legislative Research Service who, supported by and supporting the machines, can assist the member of Congress in the decision-making process.

I would apply to the Congress the existing computer capability which the military has for quick response reports on the readiness status of men and equipment.

I envision a congressional computer system which will have remote stations in every member's office and in every committee office. I should like to set as our goal a push-button system which will send back a quick response to a request — by push button — for such factual information as the status of current bills in the Congress and certain constituent information for each congressional district.

When we move beyond this limited capability into the area of computer support for decision making, we must begin with a question in a member's office or a committee office — a question put into a system but interpreted at the Legislative Research Service so that it can be put to the machine in a form that will elicit an accurate response. This is why a very important ingredient in a congressional ADP system will be a staff of highly trained interpreters in the LRS.

In talking about computers aiding the congressional decision-making process, one should keep in mind the structure of the federal government. True I have urged "separate but equal" facilities for the Executive and

Legislative Branches. But this implies an equal function for the two branches. There is no such equal function.

It is *not* up to the Congress, with the aid of ADP, to design a new bomber or a new missile. It *is* up to the Congress, hopefully with the aid of ADP, to decide whether the Executive branch made the proper decision to build the new bomber or the new missile. Congress sets the goals, approves the funds, and then checks to see that the Executive stays on the track.

Computers can render their greatest service to the Congress in the area of what might be called Advanced Decision Making.

As the complexity of our society and the means of governing it increases, the decision makers must be in a position to weigh possible options for action.

A recent management study of the Congress by the A. D. Little Company said that "Congress should develop an improved ability to test in advance the relative effectiveness of alternative courses of action . . . because effectiveness must be measured in tangible results affecting people, it cannot be measured solely in accounting terms."

In order to determine astutely and to set down alternatives, the staff of a congressional committee or of a member of the Congress ideally should possess the professional capability backed by computers and other aids, where appropriate, to identify and evaluate alternatives to the programs proposed by the Executive Branch.

It seems to me that in Advanced Decision Making we must talk in terms of "programming an environment" so that a total set of information and options is available to one who makes the decision. If this is true for the Executive, it is certainly true also for the Congress if it is to carry out its basic function of checking and balancing Executive actions and decisions.

As an example of what I mean by "programming an environment," consider the computerized model of the U.S. economy proposed recently by the American Bankers Association. The Brookings Institution supported by the ABA is building such a model.

Today if we had such a computerized economic environment we could test various alternatives. We could propose higher taxes and lower interest rates or just the opposite or infinite variations between the two until we could fairly hope we could reasonably pick the optimum combination.

But just a word of caution here which I'm sure you are already exercising in your use of ADP. Human decision making takes a little time. The decision maker should not be goaded by the machine into making premature judgments just because the machine DOES produce its results so rapidly.

Information systems are not, as Kenneth Janda points out in "Information Systems for Congress," devices for "grinding out policy decisions, and they are not designed to replace human judgment. Rather they are intended to provide . . . the Congressman . . . with knowledge for making informed choices."

But, having called for a Congressional ADP system, I would emphasize that a thorough orientation and education for all involved in such a system is mandatory.

Before any congressional system is ordered there should first be an extensive planning period. All possible alternative systems should be considered. In fact, a congressional system would have to be tailored to the particular needs of the Congress, taking into account congressional procedures, identifying the kinds of information the Congress needs, the current sources of that information, and the patterns by which it should be returned to Congress for quick use. And the congressional system should be tailored to the individual member as well as to Congress as a whole.

One idea that has been discussed, for example, is to have a profile of every member of Congress on file with the Legislative Research Service so that quick identification could be made of a particular member's special areas of interest and certain information and material could be channeled to him automatically.

The future of the United States depends on those who make its decisions. Their judgments must be based on perspective and knowledge.

In a democracy the decision making is by the people acting through their elected representatives.

The machines can aid but cannot supplant, human judgment.

At the outset I quoted John Masefield when he said:

> Man consists of body, mind and imagination.
> His body is faulty, his mind untrustworthy,
> but his imagination has made him remarkable.

Computers can aid the human body. They can aid the human mind. But nothing can supplant human imagination.

CONTENTS

Session I

Session II

Session III

Session IV

INFORMATION SYSTEM NETWORKS
Let's Profit from What We Know

ROWENA W. SWANSON
Air Force Office of Scientific Research
Arlington, Virginia

> Continuing effort to achieve technological superiority, especially in the air and in space, is not merely a matter of national prestige, scientific ambition, or economic interests; because of its military implications, such superiority is indeed a matter of vital importance to the security of the entire Free World whose leadership and defense have been thrust upon us.[1a]
>
> General J. P. MCCONNELL
> *Chief of Staff*
> *United States Air Force*

> Unless the technology is applied by designers who know what the user's criteria of effectiveness are, it is likely to be wasted.[1b]
>
> Honorable HAROLD BROWN
> *Secretary of the Air Force*

One should not, I have been told, begin a paper or a talk with disclaimers. Therefore, I should not start by saying that this paper is not a comprehensive survey of on-going information systems, nor is it an evaluation of all proposals for network designs. I have found some merit, however, in proceeding by exclusion. This enables focusing on objectives with the ever-present danger, of course, that the bare bones can be too easily seen.

The objective of this paper is to look at what has been *done toward the building of efficient and effective information systems. Although the impetus for this paper arose from my desire to underscore, to proposers of vast information networks for science and technology, some of the problems that persist at the systems level, much less the network level, information is a commodity needed by all members of society. Principles and tools that are developed to process it may ultimately benefit the housewife as well as the industrialist, the artist as well as the scientist, the schoolchild as well as the military commander. My bird's-eye viewing in this paper of a variety of systems for a variety of purposes is an attempt to preclude parochialism among those who are and will be building the systems and the networks of the future.*

Man discovers the knowledge and builds the tools that catapult him into increasingly complex states. These, in turn, impose increasing demands on his intellect and his creativity. In his work on information processing, he is homing in on methods by which he can further exploit the natural and material resources that have been given to him.

INTRODUCTION

The information problem has been characterized as a problem of abundance.[2] Machines can measure and produce more data and men can discover more variables and objects to measure than either man or machine can cope with. This paper summarizes some of the attempts that have been and are being made to systematize the behavior of both so that both can be maximally effective and continue to enlarge their spheres of activity.

The abundance of activity and the demands that systems make for a variety of talents for a variety of purposes have tended to mask common problems and common phenomena. The hardware designer, the language designer, the systems designer, and the many users too often isolate themselves in the fields of their special skills and fail to relate how their specialties should or could contribute to overall system objectives.

This paper, therefore, preliminarily reviews systems from a systems engineering viewpoint. Thereafter it considers systems that have been and are being developed for particular classes of users — scientists, managers, persons in industry and commerce, the military, and librarians. A discussion follows of proposals for integrating systems and establishing infor-