Computing in Documentation and Scholarly Research

Warren J. Haas

Summary. Production, organization, and delivery of publications of all types and in all formats are being transformed by computer and communications technology. The system of scholarly communication is, as a result, undergoing fundamental change. Whether that change will be seen as unqualified improvement depends on the way organizational and economic issues are resolved.

Computing has transformed and substantially improved bibliographic control of books and journals in little more than a decade. There has been less fundamental change in the publication and distribution of information, but the groundwork is being laid for a general restructuring of important segments of these activities as well. The pace of future development will be governed by the speed with which organizational and economic issues are resolved. Technology itself is not a constraint. The quality of the system of scholarly communication at the end of the present decade will be determined by the wisdom (as distinct from the ingenuity) with which technology is employed. Active participation by scholars from all disciplines and by senior university administrators in the process of setting objectives and means is essential.

This article focuses on scholarly communication in an academic context. The observations can generally be extrapolated to other arenas. There is some degree of concentration on research libraries, but not to the neglect of the other, equally important, system components. This is not a technology forecast; there are thousands of predictive books and articles available. Rather, it is a brief and incomplete status report on the institutionalized ways in which the product of intellectual activity is recorded, retained, and put to use.

Overview of Scholarly Communication

The goals, objectives, and processes of scholarly communication were the subject of a 1979 report by the National Enquiry into Scholarly Communication. Members of a governing board, which included representatives from publishers of scholarly books and journals, several academic disciplines, research libraries, and others concerned with the topic, sought to consider the process of scholarly communication as a whole. In the words of their report (*1*, p. xi),

> The various constituencies involved in scholarly communication, the scholars themselves, the publishers of books and learned journals, the research libraries, and the learned societies are components of a single system and are all three fundamentally dependent on each other. Moreover, this single system in all its parts is highly sensitive to influence from two outside factors—the activities of funding agencies and the development of new technologies.

While many individuals and organizations are involved, it is the set of processes designed to accomplish specific goals that constitutes the system. Figure 1 provides a compressed description. All the activities incorporated in Fig. 1 seem necessary to the system, regardless of the subject of the record or the format of the publication. A substantial volume of published information enters the system as a result of nonscholarly activities and is retained for potential scholarly use. Another body of information enters the system unpublished (manuscripts, raw data, archives, and so forth) and is retained for its research potential.

While Fig. 1 shows processes, it does not give system objectives. The objectives advanced by the National Enquiry provide a useful touchstone for assessing new directions and means (*2*):

> *Access.* Readers should have access to a comprehensive bibliographic system that allows them to identify and locate material and to obtain it at a reasonable cost without delay.
>
> *Entry.* Authors should find a variety of book publishers and journal editors willing to give a manuscript a fair reading and committed to a decision based on scholarly merit.
>
> *Quality control.* The system should have the capacity to differentiate between works of greater and lesser quality . . . and . . . importance and to match the form of publication to these differences.
>
> *Timeliness.* Manuscripts should be accepted or rejected promptly, and works should be published on schedule. Advance announcements should keep scholars apprised of forthcoming books and articles, and distribution systems should make completed work available rapidly.
>
> *Coordination.* The participants in the communications venture—scholars, publishers, technologists, scholarly societies, government and foundations, and libraries—should be mindful of their obligations and their interdependence, and pursue their goals in light of the effects their actions have on others and on the entire system.
>
> *Adaptability.* Since the needs of scholars, the tools of scholarship, the uses of knowledge, and the economic and social environment are constantly changing, the scholarly community should maintain a responsive attitude toward the elimination of obsolete methods and materials and toward possibilities of product innovation.
>
> *Financial viability.* Financing arrangements should ensure the economic viability of each function essential to the system of scholarly communication.

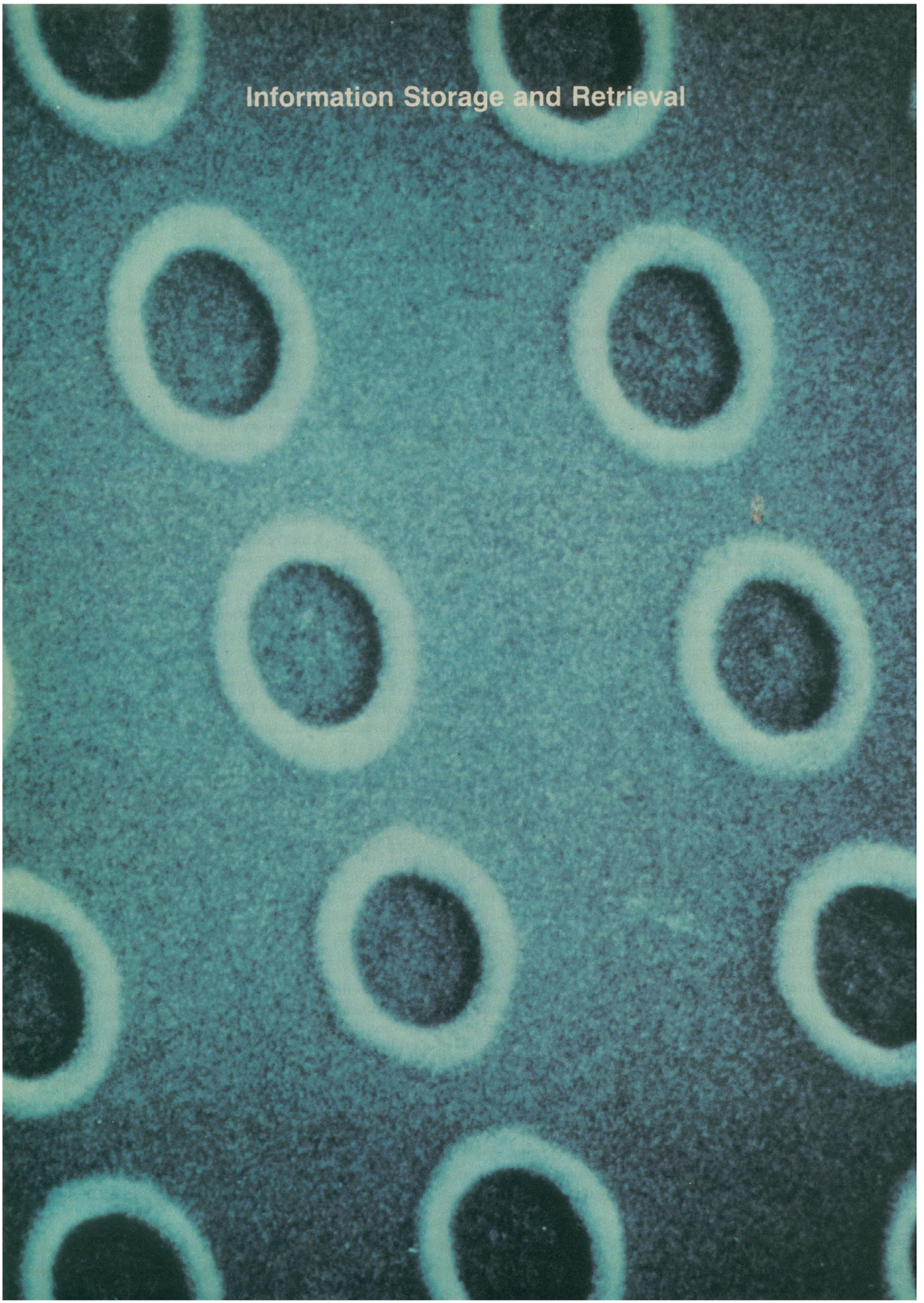
Pressures for Change

After years (some would say generations) of stability, scholarly communication is now in a period of extensive change. All organizational components and all processes are involved. The driving forces, for the most part, are not new; rather they are old problems amplified by severe economic difficulties and technological prospects.

The sheer volume of recorded information reflects both the success and failure of the system. Information about scholarly production and publication worldwide is difficult to collect and assess, but it is certain that the quantity of material entering the system is growing at an increasing rate. To give some sense of publishing volume, we note these facts: (i) according to a recent *Economist* article (*3*), "500,000 pages of technical reports, scientific journals, and books were being produced every minute" several years ago; (ii) it is estimated that more than 50 million journal articles are cited in existing machine-readable databases, most of which go back less than 10 years; and (iii) a bibliographic data-

Warren J. Haas is president, Council on Library Resources, One Dupont Circle, Washington, D.C. 20036.

0036-8075/82/0212-0857$01.00/0


Information Storage and Retrieval

base of currently published serial titles cooperatively developed and maintained by the Library of Congress, the National Library of Canada, and a small number of research libraries now includes 369,000 titles (the January–April 1981 issue of *New Serial Titles*, published by the Library of Congress as a product of that database, has 938 pages). These are signs that uncontrolled growth of published literature, especially in journals, will continue and even be accepted, implicitly, as evidence of research vitality. There is little evidence of serious effort by any of the major disciplines to judge the quality of currently published academic journals and to act on the results.

The constantly growing volume of production has had a ripple effect on all functions and organizations in the system of scholarly communication, generating dissatisfaction among users, frustration among operators, and concern about costs among administrators. Instead of controlling growth, the emphasis has been on accommodating growth, and costs of producing, storing, and processing information have escalated to levels that institutions now find hard to support. Transformation of the system, organizationally and functionally, is seen by many as essential if capabilities are to be maintained and costs controlled. The application of computer and communications technology is considered, rightly or wrongly, as the means to that end.

Computer Applications: Publishing

After 20 years of computer typesetting, computers have begun to influence other facets of publishing, but not the product of publishing itself. Text preparation, page makeup (*4*), and inventory control are all affected. More pertinent to scholars is the prospect of publishing books on demand—in essence, production of limited quantities of a document as copies are required (*5*). Already in use in specialized segments of the publishing industry, demand printing captures and prints electronically stored information by means of lasers or xerography. The end product is still print on paper, but the costs of typesetting, makeup, presswork, and warehousing are eliminated or substantially reduced. Prospects are thus improved for the production of works of scholarly importance having low sales potential and for keeping important titles available without prohibitive storage costs.

An indication of a possible new trend is found in a proposal recently advanced by a consortium of European publishers to the British Library Lending Division. The objective of the plan is to help the library satisfy requests for copies of journal articles and to convey royalties to the publishers. The projected system would encode and store journal articles on an optical videodisk for printing on demand. Articles requested of the library (which now dispatches 2 million photocopies annually) would be provided to the library by the consortium, with costs for the article itself and for related library services passed on to the user. At the least, the projected system blurs the initial distribution function of publishers and the provision function of libraries.

A more revolutionary prospect for publishing is the storage and distribution to users of text in machine-readable form. The publishing world is just beginning to explore ways to use the growing number of text databases now created in the publishing process (*6*). For example, the *New York Times* information bank has just begun to provide full text on-line, and a growing body of American and English case law is now available and is searched on-line more than 25,000 times daily (*3*).

Computer Applications: Cataloging and Indexing

In less than 20 years—with most of the progress coming in the past decade—computers have transformed book cataloging and journal indexing procedures and hence have substantially improved the process of selection. Major journals for most scientific and technical fields and for many subject areas in the humanities and social sciences are fully indexed, and the results are made available through reasonably accessible search services. Bibliographic records for a large portion of currently published monographs and serials are in machine-readable form for library cataloging purposes. The incorporation of Chinese, Japanese, and Korean vernacular script into primary bibliographic files is anticipated during 1983 (*7*).

A summary of the history of bibliographic computerization is provided in Table 1. Events noted are selected to indicate directions. The list is neither comprehensive nor necessarily balanced.

The Online Computer Library Center (OCLC) is the largest of several computerized library service systems. More than 3200 libraries are served through 22 state and regional networks. Libraries of all types participate—research, collegiate, public, corporate, and school. Over 4250 terminals are linked to the computerized database in Columbus, Ohio, through an extensive telecommunications network. The database, containing more than 8 million records, adds 25,000 records each week. The system was established as a cooperative cataloging enterprise, with individual libraries contributing new records or drawing on records in the file to catalog items being added to their own collections. Computerized records are used to print catalog cards, which are mailed to the libraries. This activity still dominates OCLC operations. Each week, more than 325,000 books are cataloged in the system and more than 2.1 million catalog cards are produced and mailed to libraries.

In recent months OCLC has added an interlibrary loan support system and an acquisitions system. Because locations are noted in the database, location and requesting procedures are simplified and delivery is expedited. The development of less expensive computing equipment with capabilities suited to bibliographic services and library operations has prompted OCLC to begin exploration of new applications, including local and regional on-line catalogs and acquisitions and circulation systems in a distributed computing environment.

Another computerized library service system is the Research Libraries Information Network of the Research Libraries Group (RLG). An institutional partnership now numbering nearly 30 research universities, RLG is a pioneering effort to address some of the fundamental problems facing research libraries and the scholarly communication process itself. Long-term programs addressing collection preservation, coordinated collection development, and extension of research support services are the substance of the RLG program. The computer-based bibliographic system, being used for shared cataloging, acquisitions, and other functions, is based at Stanford University and is viewed as the means to RLG program objectives. The system maintains the catalogs of each institutional member on-line and serves as a union catalog. The database is authority-controlled and search capabilities are extensive (names, titles, subjects, character strings, and Boolean searches, to name a few).

While automation of the bibliographic service structure is well advanced, there is still much to be done. Bibliographic systems have been developed independently and often incompatibly. Economic factors and system constraints limit their flexibility, and in the long term they need to be functionally linked. Efforts to reduce redundancy in bibliographic ef-

forts now center on creating a single name authority file to be established and maintained by a number of libraries working under direction of the Library of Congress. The file, in essence a dictionary of accepted forms for personal and corporate names along with references to variants, is a key to future cooperative efforts to establish a comprehensive, nonredundant bibliographic database, retrospectively as well as prospectively. The Library of Congress and several bibliographic services are well along in systems design and telecommunications planning work to establish the computer-to-computer links required to implement authority file operations and, ultimately, a number of other library applications.

There has been less progress in providing subject access to monographic literature, especially in the humanities and social sciences, and new analytical studies of needs and methods are just beginning. The growing body of published material will demand more precision in the search process. Systematic organization of effort and better understanding of the use and structure of subject terminology seem necessary for further improvement.

Paralleling the development of comprehensive databases is work now under way to establish on-line access to machine-readable catalogs directly by users. A small number of such catalogs—the likely successors to card catalogs in many situations—have been established in academic and public libraries, and an intensive study of catalog performance, user expectations and habits, and system impact is being done. The objective is to establish guidelines for all aspects of on-line catalog design in order to promote economy of operation and fully adequate performance and to avoid unnecessary variations.

Continued leadership by such agencies as the National Library of Medicine, which has a record of moving to new frontiers in advance of the pack, and the Library of Congress, which is the principal source of primary bibliographic records and the center for both quality control and coordinated research library effort, is essential if the benefits of recent and present work are to be fully realized. Cooperation among all system components is imperative if diverse contributions are to coalesce into a cohesive system. The Bibliographic Service Development Program of the Council on Library Resources is specifically directed to these ends, providing funds and promoting cooperative development of additional capabilities. The council's program is guided by an agenda of bibliographic targets, projected for completion by 1984 (*8*).

Given the inherent complexities of comprehensive bibliographic control of recorded information, the rapid expansion of computerization, and the diversity of the participants, the world of scholarship has been well served, not only by the depth and reliability of coverage but also by the fact that the products of the new systems, from on-line catalogs to bibliographic publications, tend to equalize access to information about information. In effect, the bibliographic structure has become library-independent and has thus created a new set of concepts and service loads on components of the system responsible for delivery.

Computer Applications: Provision of Publications

The extensive computerization of cataloging and indexing activities has raised the prospect of growing frustration, at least in the short term, for individuals who now quite easily learn of the existence of items containing information but who can less readily obtain the items themselves. With the exception of experimental systems (such as INTREX and subsequent work at Massachusetts Institute of Technology) and applications in narrow fields or special environments, computers have had little effect on the processes of locating, acquiring, and delivering to users material selected from the published store.

In general, location systems presently in use by libraries are characterized by voluntary reporting of holdings to union catalogs. As those catalogs are computerized, both reporting and query are simplified, but deletion of records no longer valid is sporadic. Development of a strategy for the purposeful reporting of holdings by region or some other partition is needed.

The format of items retrieved from storage and delivered to users is largely determined by the form of storage. Either the item itself is loaned or a copy is made for use, an option governed by costs, security or preservation considerations, copyright constraints, and level of demand.

A number of studies have proposed ideal systems for storage, delivery, and use [for example, full-sized text for use, photographically reduced text for storage, and digitized text for transmission (*9*)]. It is evident that the technology is already available or soon will be, but significant change is unlikely until demand is established, questions of cost and funding resolved, and legal issues settled (*10*).

Unresolved Issues

Information is both a public good and a commodity. It is used but not consumed. Its utility changes with time, but in no fixed pattern. Its value is measured in intellectual terms and in economic terms, but there is no necessary correlation between the two. Control of information is power, but unconstrained access is a public goal. Computerization of the processes of producing, organizing,

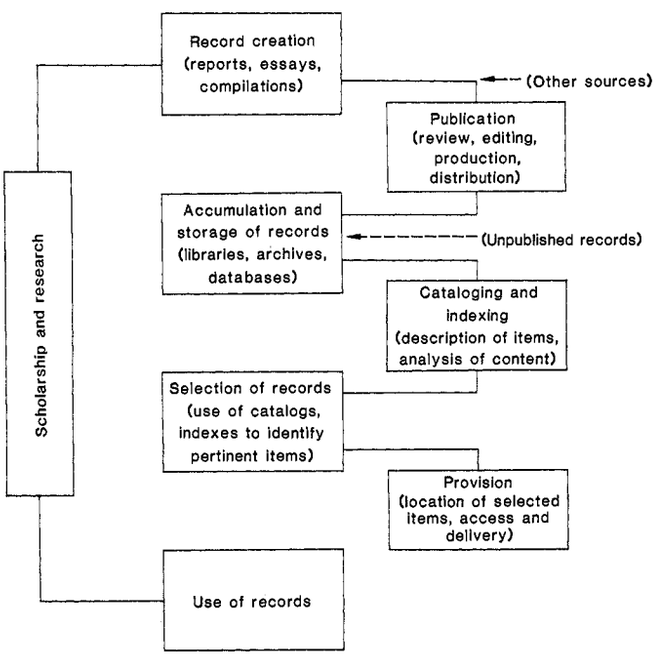


Fig. 1. The scholarly communication system.

and providing information forces the recasting of long-established relationships, changes control and use patterns, introduces new costs, and confuses the "intellectual audit trail" inherent in peer review and publication in a fixed and tangible form.

These and other characteristics pose many questions. The answers will have much to do with both system automation and the substance of scholarly communication. Among the key questions are the role of government in the resolution of commercial and legal issues, the funding of old and new costs, especially in research universities, and the quality of participation by scholars in policy formulation.

Government is a major producer of information, and government-supported research is a major consumer, most heavily in scientific, medical, technical, and quantitative social science fields. The level of use of existing documentation systems and the development of new computer-based systems and services will be governed in part by the nature of the relationship between the public and private sectors. Formulation of policy in this arena will determine the manner in which information created with public funds will be distributed and will substantially affect the service roles of national libraries, the Government Printing Office, and every department and agency.

A recent report by a task force of the National Commission on Libraries and Information Science (*11*) discussed key issues and set forth the views of two opposing sides: those who would not restrict the role of government with respect to providing information products and services and those who would limit government activity and rely instead on market forces. While computer and communications technology are not considered specifically, the prospect of new approaches to marketing information inherent in that technology raised the issue in the first place and permeates the general thrust of the report.

The task force did not resolve the fundamental conflict, but did agree on a number of principles and recommendations designed to provide a cohesive base for the evolution of information services. While it is impossible to report even the substance of the 27 recommendations, the underlying principles are as follows (*11*):

1) The federal government should take a leadership role in creating a framework that would facilitate the development and foster the use of information products and services.

2) The federal government should establish and enforce policies and procedures that encourage, and do not discourage, investment by the private sector in the development and use of information products and services.

3) The federal government should not provide information products and services in commerce except when there are compelling reasons to do so, and then only when it protects the private sector's every opportunity to assume the function(s) commercially.

4) The federal government, when it uses, reproduces, or distributes information available from the private sector as part of an information resource, product, or service, must assure that the property rights of the private sector sources are adequately protected.

5) The federal government should make governmentally distributable information openly available in readily reproducible form, without any constraints on subsequent use.

6) The federal government should set pricing policies for distributing information products or services that reflect the true cost of access and/or reproduction, any specific prices to be subject to review by an independent authority.

7) The federal government should actively use existing mechanisms, such as the libraries of the country, as primary channels for making governmentally distributable information available to the public.

The task force did not consider problems of international data flow, a matter of growing importance. There is some evidence that regulations, especially in Europe, might constrain the international flow of computerized bibliographic information. A study sponsored by the International Federation of Library Associations and Institutions is now under way to assess the problem, which may be even broader. Ithiel de Sola Pool (*12*) asserts that "trade restrictions on transborder data flows are not merely trade

Table 1. Computerization of bibliographic processes: highlights since 1961 (*14*).

Date	Events
1961	Development of Medical Literature Analysis and Retrieval System (MEDLARS), the basis for the first automated library bibliographic network. In batch mode, the system permitted rapid printing of *Index Medicus*, the bibliography of current medical literature.
1961	Institute for Scientific Information and *Science Citation Index* established. This and subsequently developed indexes for other areas use citation frequency to organize literature and to map research progression in specific fields.
1965	Information Transfer Experiment (INTREX) conducted at MIT to store, retrieve, and transmit scientific information electronically.
1965	Conference of Library of Congress staff and academic librarians to explore bibliographic automation. One result was development of Machine-Readable Cataloging (MARC), now the standard communications format for bibliographic information.
1967	Ohio College Library Center created. OCLC evolved to become the largest bibliographic service system, supporting shared cataloging, interlibrary loan, and acquisitions.
1968 to 1974	Rapid growth of commercial on-line search services, including DIALOG, ORBIT, and Bibliographic Retrieval Service (BRS).
1970	University of Chicago Library development of comprehensive Library Data Management System to computerize and integrate primary library functions.
1972	MEDLARS converted from batch to on-line system, MEDLINE.
1972	Bibliographic utility formed at University of Toronto. Shared cataloging and other services were gradually developed and extended to Canadian libraries and, in 1980, to the United States.
1973	Conversion of Serials (CONSER) project begun, the first full-scale effort to build a machine-readable database of records contributed by a group of cooperating libraries.
1973	Research Libraries Group formed. The organization now operates the Research Libraries Information Network bibliographic service and supports cooperative programs in collection development and preservation.
1973	Universal Bibliographic Control (UBC) established. This continuing program of the International Federation of Library Associations and Institutions promotes development of national bibliographies and international exchange of standardized bibliographic records.
1974	BALLOTS, an on-line technical processing service, developed by the Stanford University Libraries.
1976	Washington Library Network (WLN) established. The Washington state legislature supported development of the first authority-controlled bibliographic network, now used by libraries of all types in the Pacific Northwest.
1978	Bibliographic Service Development Program established by the Council on Library Resources. With the goal of establishing a nationwide bibliographic service, the program promotes cooperation among bibliographic services of all kinds among academic and research libraries. It funds and manages projects, including on-line catalog development, creation of name authority file service, and functional linking of existing computerized services.
1981	Library of Congress "freezes" its card catalog. Records for items added to Library of Congress collections after 1 January 1981 are added only to computerized bibliographic systems.

restrictions. They are restricting our freedom of speech and cultural relations. If we allow that kind of regulation to be treated as just one more nuisance about which we have to compromise, we will have lost a much bigger game."

Copyright issues are closely related to national information policy and regulatory control of communications. In a recent speech, Register of Copyrights David Ladd called for legislative action to protect the rights of authors threatened by new technologies. On the same theme, Dan Lacy (*13*) notes: "As textual materials come to be used in systems rather than reproduced in copies, we need to develop copyright methods analogous to those already existing in music and drama, where the creator's reward is related primarily to the performance (that is, the user) rather than the copies of the work."

Costs of computerizing information services, institutionally and nationally, are great. Initial investment requirements are substantial and come at a time when the scholarly world and research enterprises are hard pressed. Traditional publishing, despite efforts at cost containment, grows more expensive each year. Support for drastic changes to substantially alter costs (for example, development of new format options) is not yet strong enough to encourage decisive action. Computerized operations and services, while sometimes cost-effective individually, represent new costs in the aggregate, in part because they often complement rather than replace existing services and in part because communications open a new cost category and imply a new pricing structure for information itself. Research libraries, publishers of scholarly books and journals, and, ultimately, research universities are faced with a funding problem of great magnitude, one that has its roots in the transformation of the system of scholarly communication. There is a special irony here, since universities are the home of much research and scholarship and thus the source of much of the record that enters the system. That they must, collectively, buy back what they produce individually at a cost that endangers their capacity to exist, or at least thrive, is perhaps the most intractable of our underlying dilemmas.

Public policy and volatile economics are factors that will affect the manner in which computing is employed in scholarship and the rate at which the technology transformation is made. Another set of issues, largely personal and intellectual, also needs attention. The requirements of academic research and scholarship, because they are diffuse and not because they are unimportant, will not determine to any significant degree the future technology of information services. The scholarly world does have an obligation to see that the institutions and organizations it values are not jeopardized by the competition for funding that is implicit in the introduction of new and costly systems that, despite strong assertions to the contrary, will probably supplement rather than supplant printed books, scholarly journals, and research libraries.

In an epilogue to the first chapter of *Scholarly Communication* (*1*), Herbert Bailey, director of the Princeton University Press, described the life of a working scholar after computer and communications technologies have been integrated into publishing and libraries:

> The boundaries between scholar-authors and publishers and the book trade . . . and libraries and scholar-readers will shift and blur. New library-like services will be offered by publishers and wholesalers, scholars will enter material directly into libraries, libraries will perform publisher-like and bookstore-like functions. The new national and ultimately international information system will eventually be established. . . . The problems in establishing such a system are not mainly technical; they are organizational. . . . But in the long term, the driving force for change will be economic because the present system is reaching a state of saturation that is increasingly unworkable and costly.

Publishers, librarians, and scholars in all disciplines are reflecting on the prospect of merging new systems with the old. There is little support for the extreme position that established components of the scholarly communication system will be replaced by the new technology. There is concern that the resources and skills required to transform the old and wisely use the new will be too little and too late.

References and Notes

1. *Scholarly Communication: The Report of the National Enquiry* (Johns Hopkins Press, Baltimore, 1979).
2. *ibid.*, p. 7.
3. *Economist* **281** (No. 7206), 103 (1981).
4. R. Shotwell, *Publishers Weekly* **219** (No. 15), 34 (1981).
5. ———, *ibid.* **219** (No. 7), 47 (1981); *ibid.* **219** (No. 23), 50 (1981).
6. R. Dahlin, *Electronics and Publishing* (Bowker, New York, 1981).
7. *Automation, Cooperation and Scholarship: East Asian Libraries in the 1980's* (American Council of Learned Societies, New York, 1981), p. 6.
8. Council on Library Resources, *Bibliographic Targets for 1984* (Council on Library Resources, Washington, D.C., 1981).
9. M. V. Mathews and W. S. Brown, in *On Research Libraries: Statement and Recommendations of the Committee on Research Libraries of the American Council of Learned Societies* (MIT Press, Cambridge, Mass., 1967).
10. C. Rochell, Ed., *Information Agenda for the 1980's: Proceedings* (American Library Association, Chicago, 1981); in particular, see L. M. Branscomb, pp. 43–53.
11. *Public Sector/Private Sector Interaction in Providing Information Services* (Comm. Doc. 81-34, National Commission on Libraries and Information Science, Washington, D.C., 1981).
12. I. de Sola Pool, remarks at the International Council on the Future of the University, New York (1981).
13. D. Lacy, in (*10*), p. 105–111.
14. Table 1 was compiled from the following sources: Council on Library Resources, *20th Annual Report* (Council on Library Resources, Washington, D.C., 1976); A. T. Hamlin, *The University Library in the United States* (Univ. of Pennsylvania Press, Philadelphia, 1981); F. W. Lancaster, *Toward Paperless Information Systems* (Academic Press, New York, 1978); W. J. Welsh, in *Academic Libraries by the Year 2000: Essays Honoring Jerrold Orne*, H. Poole, Ed. (Bowker, New York, 1977), pp. 177–190.