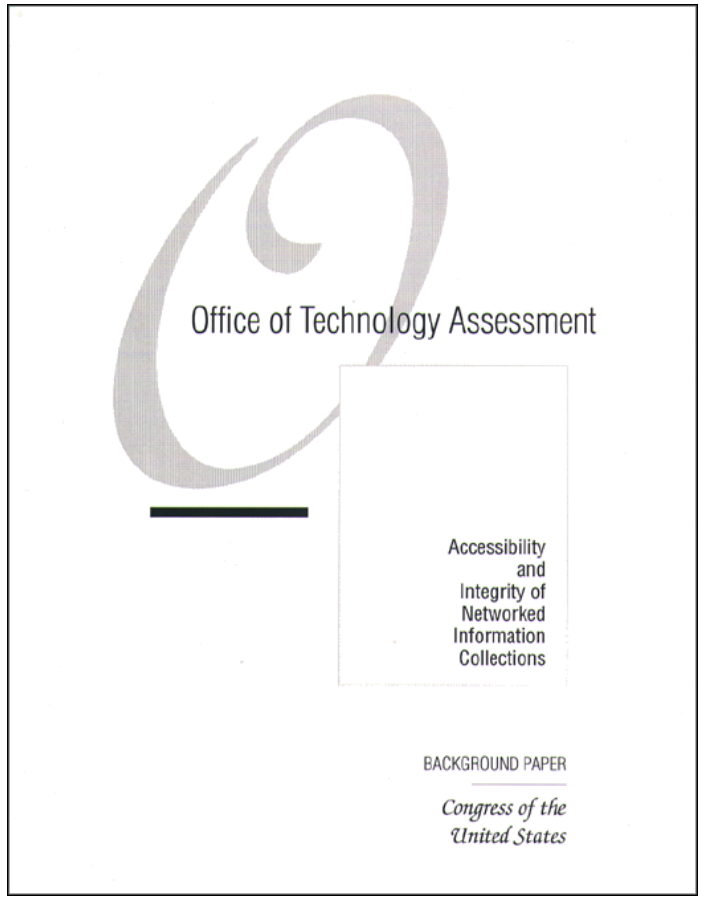
Accessibility and Integrity of Networked Information Collections

August 1993

OTA-BP-TCT-109
NTIS order #PB93-218923





Foreword

Technological advances in networking--ranging from widespread use of Internet to development of the National Research and Education Network and National Information Infrastructure--are ushering in a new era of information systems and online information resources. Networked information collections, or "digital libraries, " allowing online access to books, journals, music, images, databases, and multimedia works will be an important part of this new infrastructure. The extent to which these resources' potential is realized will depend, in part, on their accessibility and "user-friendliness. " But measures to increase user friendliness have an equally important counterpart in measures to ensure the integrity and proper use of network information resources. Otherwise, potential problems like plagiarism, corruption of databases, and improper use of copyrighted materials could impede development of networked information collections or limit their public accessibility.

Clifford Lynch's contractor report was prepared as part of an Office of Technology Assessment planning activity on the accessibility and integrity of digital libraries, in partial fulfillment of Section 2385 of the Rural Economic Development Act of 1990 (Public Law 101 -624).

In this contractor report, Lynch takes the perspective of the library community and its users and patrons in examining issues and potential problems facing digital libraries, with emphasis on accessibility, integrity, and the interrelationships between them. Lynch discusses the technological and institutional measures that can be used to address access and integrity issues, identifies problems that cannot be appropriately resolved by current technologies and/or institutions, and offers his views concerning actions by Government and others that will be needed to address them.



Roger C. Herdman, Director

REVIEWERS

Robert J. Aiken
U.S. Department of Energy

Pat Battin
Commission on Preservation and Access

Ann Bishop
University of Illinois

Bonnie C. Carroll
Information International Associates Inc.

Steve Cisler
Apple Computer, Inc.

Nancy Eaton
lowa State University

Jane C. Ginsburg
Columbia Law School

Jane Bortnick Griffith
Congressional Research Service
Library of Congress

Lee A. Hollaar
University of Utah

Karen Hunter
Elsevier Science Publishing Company, Inc.

Brian Kahin
Harvard University

Brewster Kahle
WAIS, Inc.

Tim Lance
State University of New York at Albany

Jean Pony
NYSERNet

Maria Zemankova
National Science Foundation

OTA appreciates and is grateful for the valuable assistance and thoughtful critiques provided by these reviewers. They do not, however, necessarily approve, disapprove, or endorse this contractor report.

ACCESSIBILITY AND INTEGRITY

OF

NETWORKED INFORMATION COLLECTIONS

by

CLIFFORD A. LYNCH

Contractor Report prepared for the
Office of Technology Assessment
Telecommunication and Computing Technologies Program
Joan D. Winston, Contract Coordinator

The views expressed in this contractor report are those of the author and do not necessarily reflect the analytical findings of OTA, the Technology Assessment Board, or individual members thereof.

Accessibility and Integrity of Networked Information Collections
Clifford A. Lynch
July 5, 1993

Table of Contents

1. Introduction 1
Objectives and Purpose of This Paper . . . . . .
Overview of the Paper. 3
Acknowledgments and Disclaimers 6
2. The Networked Information Contetex 7
3. An Overview of Access and Integrity Issues in the Networked Information Environment. 14
4. Access to Electronic information: the Changing Legal Framework 17
From Copyright and Sale to Contract Law 17
Licensing and the Interlibrary Loan System in the United States 19
Other implications of the Contract Law Frameworkfor Libraries 22
Libraries and Multimedia: New Legal issues 25
5. The Role of Secondary Information Sources and Automatic lndexing in Access to Electronic Information. 28
6. Access to and integrity of the Historical and Scholarly Record 36
7. The lmpact of Micropublishing, Narrowcasting, and lnformation Feeds. 4l
8. Privacy issues in access in the networked information environment 46
Confidentiality and Anonymity of Access 46
Who Owns Access Histories? : Privacy and Market Research 52
Privacy, Intellectual Property and Electronic Mail Enabled Communication 55
9. The Internet Tradition of''Free''lnformation: Quality, integrity and Liability Issues 57
10. Access to Electronic Artifacts: The Problems of Versions, Presentation and Standards 60
The Nature of Electronic Works 60
Presentation and Content Standards 61
The Problems of Structured lnfomation 65
11. Digital images and the integrity of the Visual Record 66
12. Authenticating versions and sources 67
13. Citing, identifying and Describing Networked Information Resources 73
14. Directories and Catalogs of Networked information Resources 78
15. Conclusions 81
Integrity and Access Issues in the Broader Networked Information Context 81
Ensuring Access to Information in the Networked Information Environment 84
Privacy, Confidentiality and Anonymity in Access to Electronic Information 86
Infrastructure and Standards. 87
16. Recommendations for Possible Action 91
Legislative, Government and Public Policy Actions 91

Actions by the Stakeholders: Authors, Publishers, Libraries, Information Technology Providers and the Education Community 93

Glossay 94
Suggested Background Readings 99
Networking Technology, Networks, and the Internet 99
Networked Information Services and Networked information 99
Surveys of Topics Related to this Paper 99
References 101

Accessibility and Integrity of Networked Information Collections

July 5, 1993

Clifford A. Lynch

This is an Office of Technology Assessment outside contractor report. It does not necessarily reflect the views or findings of the Office of Technology Assessment or the Technology Assessment Board.

1. Introduction

Objectives and Purpose of This Paper

We are entering an age when a great deal of information is available in electronic formats and can be obtained through computer communications networks. Sometimes, such collections of network-accessible electronic information are referred to as “digital libraries” (although, as discussed later, I view this terminology as somewhat misleading; indeed, one of the issues explored here is the developing roles of such electronic information collections and their relationships to institutions such as libraries). These new electronic information resources are still being defined and almost everything about them is poorly understood; conventions, expectations, social customs, institutional roles in society, business relationships, legal precedents and public policy choices based on print and mass media broadcast environments may no longer be directly applicable in the networked information environment. This new environment challenges us to reconsider assumptions and to rethink many aspects of the current systems of information creation, management, organization, preservation and use, as well as the broader public policy context of these systems.

This paper, prepared under contract to the United States Congress Office of Technology Assessment, is intended to consider questions related to the integrity and accessibility of these new electronic information resources. These are key areas that pervade many of our assumptions about all forms of information; the resolution of integrity and access questions in the electronic information environment will play a major role in defining how we create, use and manage these new forms of information and the economic system that will support these activities, the purposes it will serve within our society, and the functions that various institutions will need to carry out in this new environment. Both integrity and access are complex issues, and have intricate

relationships not only to each other but to legal and public policy considerations such as privacy and intellectual property. Further, integrity and access cannot be considered outside of the context of the communities and institutions that create, distribute, organize, manage and use information; this leads to a discussion of the evolving roles of libraries, publishers, and the authors and readers of various types of material. Finally, because the growing flood of available information is increasingly overwhelming and unmanageable without mechanisms and conventions to allow us to identify, select, and cite particular works, I have devoted considerable space to exploring how these mechanisms are developing in the world of electronic information resources.

While there is some emphasis here on printed information and its electronic descendants, and particularly scholarly communication and publication, i have tried to at least begin an examination of the broader issues related to mass communications, newspapers and news broadcasts, government information, and new multimedia information resources that might be of interest to the general public rather than just the research and higher education communities. The reader should recognize that my comments in these areas are at times rather speculative; the research and higher education communities have been early adopters of networked information, and while even within those communities the issues are far from resolved, we have much more experience with scholarly information than we do with mass market network-accessible information.

To a great extent, this paper presents a personal view of the developing issues. My background is large research libraries and the higher education community. Other stakeholders in the areas discussed here, such as publishers or public librarians, may well have other views, and certainly even within my own communities there are many different opinions not only on what should be done but even the significance of many of the trends and developments discussed here. While I have attempted in various places in this paper to at least indicate the existence of other viewpoints and sometimes to sketch the rationale behind them, the reader should not view this paper as a fully balanced survey and synthesis of the various viewpoints on the issues. Finally, I must stress that I am not an attorney, and thus my comments on legal issues should be viewed as probably better informed about the implications of current legislation and case law on systems of information creation, dissemination, preservation and use than on the legal specifics themselves. While this paper has benefited greatly from reviewers more learned in the law, I may well have misunderstood or overlooked specific legal issues despite their efforts to help.

I have not provided an explicit executive summary of this paper. Those interested in obtaining a quick overview of the paper's coverage and conclusions should read the subsection directly following this one which gives an overview of the paper, and then proceed to Sections 15 and 16, the conclusions and recommendations for possible follow-on actions. These proposals are somewhat limited in nature; my purpose in this paper is primarily to illuminate the relationships among technological developments, electronic information, legislation and public policy, and various institutions such as libraries, and to identify and define the issues that new technologies are bringing to the fore. In a number of areas I have suggested consideration of a review and reexamination of current public policies, legislative positions, and government and private sector investment decisions in light of the trends described here; but in cases of public policy and legislation I have focused on providing information that could help to

inform such a review and reexamination rather than attempting to suggest specific outcomes. I have mentioned a few specific areas where I felt that there was a need for funding, to move the implementation of existing policy directions.

Overview of the Paper

The paper begins with a survey of recent developments in networked information resources and tools to identify, navigate, and use such resources (Section 2). This is a vital context and source of examples for the remainder of the report. As these new information resources are surveyed, this section also examines the idea of "digital libraries" and the relationship between today's libraries as organizations and collections of electronic information. Those readers unfamiliar with the rather extensive and quickly evolving developments in networked information resources may find this section of the paper particularly difficult reading, heavy with acronyms and references to a bewildering array of organizations and systems. To help those readers, I have provided a brief glossary of terms that I felt might be unfamiliar and a short annotated bibliography of useful background readings (as distinct from specific references cited in the text).

Building on the context established in Section 2, Section 3 provides an overview of the issues involved in access and integrity questions and the relationships among them. The boundaries of the paper are also largely defined here: for example, issues of network access (as opposed to access to information on the network) are excluded, as are most of the public policy issues specific to government mandates to provide access to government information. If access is defined somewhat narrowly, integrity is treated quite broadly and encompasses not only conventions and audit trails needed to ensure consistency and accountability in the scholarly, historical and cultural records, but also questions of content bias and accuracy. The links between access and integrity are stressed: for example, ensuring the integrity of a corpus of publication is meaningless if there is no access to this body of information.

The paper then discusses the changing legal framework that governs use of electronic information as contract law rather than simple sale within the context of copyright law becomes the dominant model for acquiring access to electronic information (Section 4). This shift is shown to have profound implications for access to information and also for the integrity of the historical, scholarly and cultural record that "published" information has traditionally represented. The effects of this shift on libraries and the interlibrary loan system, which has historically been a key part of the library community's strategy for providing their patrons with very broad access to information, is examined in depth. This is followed by an exploration of the new and even more complex questions raised by image and multimedia content in electronic networks; here we start from an environment of much greater ambiguity when applying copyright law, and find both libraries and rightsholders facing substantial problems in understanding how to manage these materials. Because of this, we find that the shift to contract **law** offers stronger operational benefits for those institutions (including libraries) that want to acquire access to image and multimedia information, although this shift again raises serious public policy issues.

The role of secondary information sources in providing access to, and evaluation of, electronic information is examined from several perspectives in Section **5;** these include the role of the extensive and well-developed marketplace that exists today in databases (for example, so-called abstracting and indexing databases) and other tools that provide access to the print literature (and in future to electronic source material) and the potential impact of new tools derived from technologies such as automatic indexing. Appropriate application scenarios for the different access tools are discussed, as is the growing power of these access tools in defining the part of the "published literature" that will be read and evaluated by various communities. This power places great responsibility for quality and integrity on the secondary information providers and access tools and thus plays a significant role in establishing the overall accessibility and integrity of the published literature. Considerable attention is given to the implications and impact of quality and integrity problems in this area.

The paper then turns to the central issues of the historical and scholarly record and the access and integrity questions that surround it as much of the information that comprises this record shifts to electronic forms (Section 6). Much of the theme here revolves around the multiple uses to which this record is put, the different expectations and requirements of the various communities that rely on this record, and the social, business, legal and technical issues involved in trying to address these somewhat conflicting needs. The implications of the shift from sale under copyright to contract law based licensing emerge clearly here as a potentially serious threat to both access and integrity for this record, and help to frame the public policy issues involved in ensuring its preservation.

A series of technology developments, economic factors and market demands have led to the creation of ever more specialized publications; in the print world this is often termed "micropublishing" while in the broadcast industries it is sometimes referred to as "narrowcasting. " The trend towards increasingly personalized information products, both in print and digital forms, combined with the appearance of new electronic information sources such as news feeds or sensor data streams that can essentially be "personalized" by the recipient (or the recipient's organization) using digital technology again raise serious challenges to the accessibility and integrity of the electronic record. They create enormous problems for libraries as the traditional managers of this record. And they raise new, complex public policy problems related to equality of access to information sources, and the implications of inequitable access by various sectors of the population. Section 7 explores this area.

The relationships between privacy and access in the electronic environment are very complex. Do users have a right to anonymous access to information resources, or at least a reasonable expectation that records of their information seeking and use should be kept confidential? Complicating these questions are conflicting requirements for accounting, cost recovery, system management and tuning, ensuring the security and integrity of the underlying computer networks, and the economic motivations of various parties to collect usage and demographic data for potential resale or reuse that run counter to the long standing policies of confidentiality and anonymity that have been established within the library community. At the same time, technological developments are in some cases preempting (at least in the near term) the as yet unresolved policy debates. Section 8, on privacy and access attempts to summarize the issues in these areas and to reflect some of the competing pressures.

Much of the information currently available through computer networks is “free”; that is, the user is not charged for using it. Section 9 examines some of the implications of free information, such as expectations about accuracy, timeliness, and value. In a real sense, the extensive reliance upon free information sources is shown to add a certain instability to the networked information environment because of the ease with which inaccurate information can quickly spread with little accountability. While the public policy questions here seem to be rather limited, this is important to developing a full picture of the networked information environment.

The paper then considers the nature of electronic works and how access to such works can be provided in Section 10. Here, one major theme is the current tendency to intertwine content and access mechanism. This has serious implications for long-term access and preservation of these works as the information technology environment continues to change. There are also subtle integrity issues that arise as we attempt to define the distinctions between content and presentation of that content. This section also emphasizes the importance of establishing and encouraging the widespread adoption of appropriate standards that allow information providers and users to separate content, presentation, and access or viewing tools for electronic works.

While we realize intellectually that photographs can be altered, the visual evidence provided by photography has provided a very important part of the historical record in our society. We have made very strong intuitive assumptions about the integrity of photography. The section on digital imaging and the integrity of the visual record (Section 11) summarizes how the development of new digital imaging technologies calls this role into question and places much greater emphasis not on the image as artifact, but on the use of verified true versions of images combined with formal, auditable links from that image to the individual or organization that has created it. This serves as motivation for the Section 12, which deals with the authentication or verification of electronic works and their creators. This requirement is a cornerstone of the integrity of electronic information; while perhaps most visible in the context of images, it pervades the use of all types of electronic information. A discussion of the issues involved in making such authentication possible leads directly to a series of issues concerning cryptographic technology, including standards, intellectual property rights and export controls (which in turn are related to the search for an appropriate balance between national security concerns and the needs for privacy and authentication in the networked environment).

The final sections of the paper consider two related issues that are again central to both access and integrity in the electronic information environment. The first, covered in Section 13, has to do with identifying and citing electronic works, and summarizes requirements, architectural approaches and standards developments to address these needs. The second issue is the intellectual identification of networked information resources; here the development of catalogs and directories for these resources is considered, with some emphasis on the role of libraries on the one hand in creating these directories and catalogs and, conversely, the way in which such directories and catalogs will integrate with existing tools used by libraries to provide access to the print literature during the long transitional period where both electronic and print information are essential parts of the scholarly communication system and co-exist. Section 14 addresses these questions.

The paper divides its conclusions into two parts The first concluding section (Section 15) tries to summarize and tie together the various developments, trends and issues that have been surveyed, and also to set the conclusions of this paper in a broader context. One point that is emphasized in this section is that two of the key areas slowing progress in infrastructure development for networked information—standards development and the deployment of integrity and authentication support services based on cryptographic technologies—call for issues well outside the primary scope of this report to be addressed, but until these issues are addressed, will continue to cause problems in our ability to resolve access and integrity questions related to networked information. Section 16 builds on and Section 15 and enumerates specific issues and projects where actions—by Congress, by various government agencies, or by various groups within the stakeholder community—should, in my view, be considered.

Acknowledgments and Disclaimers

This paper has benefited greatly from the careful work of a large number of reviewers some known to me and others anonymous. Their comments greatly improved the quality and accuracy of the work. I thank them for their efforts and their thoughts, while taking all responsibility for errors and for opinions with which they may well disagree. I owe particular thanks to Joan Winston of the Office of Technology Assessment for multiple reasons: for inviting me to explore these extremely interesting issues in the first place; for a very careful and thoughtful review of a draft of the paper; and for her patience as deadlines slipped and the paper continued to grow and evolve at its own pace towards completion. Conversations with any number of colleagues, both in person and electronically, have been of immeasurable help in understanding these issues. Particular thanks are due Steve Cisler, Michael Buckland and Howard Besser. Various electronic newsgroups and discussion lists have also been very valuable to me in preparing this paper, both in terms of discussions (where typically I listened much more than I contributed) and by providing direct object lessons about several of the topics discussed in the paper. I would also like to thank Cecilia Preston both for a careful review of a draft of this paper and for her help with the citations and references.

References and citations for a paper of this type are a huge problem. I have not even tried to be comprehensive in the citations; while a comprehensive bibliography of developments in networked information would be extremely useful, compiling it would be a truly enormous effort, and, given the rate of developments in this area, it would need continual updating. Worse, a number of the projects have not yet been much described in the published literature, and what material exists is scattered across obscure conference and workshop proceedings and technical reports. Thus, I have had to satisfy myself with the inclusion of a limited number of references suggestive of the type of work going on today, particularly in Section 2 on the networked information context, and I would beg the indulgence of those authors who have made important contributions to the literature in this area but do not find their efforts reflected in the citations here. Some readers may find the bibliography in [Lynch & Preston, 1990] a helpful supplementary source. In sections this paper touches on other large and complex areas, such as the technology of cryptography; for basic citations in this area (as well as a very readable overview of some of the technology) the reader should see [U.S. Congress Office of Technology Assessment, 1987]. Similarly, I have not provided

much coverage of the literature on intellectual property and copyright issues; an excellent source for these is [U.S. Congress Office of Technology Assessment, 1986]

The opinions expressed here are mine and do not necessarily reflect the views of any other organization or individual. A number of products and projects are mentioned here as examples; these mentions should not be interpreted as endorsements of products. Finally, a large number of trademarks and service marks are mentioned in the text. In particular: UNIX is a registered trademark of Bell Labs; Macintosh is registered to Apple Computers, as is Quicktime. MELVYL is registered to the Regents of the University of California. PostScript, SuperATM and Acrobat are trademarks of Adobe. I apologize in advance for those that I have failed to mention here.

2. The Networked Information Context

As use of the Internet becomes more widespread, it is becoming clear that access to information resources through the network will become one of the central applications of the network. The term “access,” here and *throughout this* paper, is *used in the broadest sense: not simply electronic connectivity to information resource providers through the network, but the ability for people to successfully locate, retrieve and use the information contained within various computer systems.* For a large segment of the Internet user community, such access to information and network-based communications tools such as electronic mail will increasingly be the applications which initially attract users to the Internet and subsequently are most heavily used. Further, we are already seeing both the development of a series of information discovery and navigation tools aimed at the end user and the evolution of a set of services that combine the access and communications capabilities offered by the network into new hybrid information services that are more interactive and immediate than traditional print or broadcast media. Libraries, publishers, and government organizations that, create, collect and provide access to information, are all striving to define and understand their possible new roles in this developing networked information environment.

Discussions about the development of the Internet and the evolution of the National Research and Education Network (NREN) program are increasingly taking a broader view that emphasizes not only the continued expansion and upgrading of the technical communications infrastructure but also the need for policies and planning to foster the development of networked information resources and the tools and educational programs needed to allow network users to create, navigate, and utilize such resources. Another component of this shift of focus is the recognition of the need to transform existing printed and electronic information resources into network-accessible forms. As we look beyond the NREN program, which is targeted to serve the research, education, library and government communities (broadly defined) towards the discussions about a full National Information Infrastructure (Nil) it is clear that electronic information content on future networks will play an increasingly dominant and driving role in network evolution.

This shifting focus is evident, for example, in a wide range of bills before the 1993 Congress, including S4, The National Competitiveness Act of 1993 (Hollings), S626, The Electronic Library Act of 1993 (Kerrey), HR1757, The High Performance

Computing and High Speed Networking Applications Act of 1993 (Boucher) and HR1328, The Government Printing Office Electronic Information Access Enhancement Act of 1993,[1] to name some of the most major. Various federal agency based programs (for example, at NASA, the National Agriculture Library, the Library of Congress, and the National Library of Medicine) are also underway to foster the availability of networked information resources.[2] The recent revision of the Office of Management and Budget Circular A-130 and its associated guidance to government organizations also speaks to the need to make information available in electronic form. Paralleling these activities at the federal level is a growing interest on the part of some state governments in the potential of the network to provide improved access to state, regional and local information (for example, in Texas [Stout, 1992]. Colorado, Utah,[3] California and North Carolina). State libraries are using the Internet as a focus for statewide multitype library planning in several states, including Colorado [Mitchell & Saunders, 1991] and Maryland.[5]

Broader based initiatives that span the government, research and education, and commercial sectors recognize networked access to information resources as a key element. For example, the National Science Foundation (NSF) sponsored Synthesis Coalition, which is focused on improving engineering education at all levels, includes a component called NEEDS (the National Engineering Education Delivery System) which focuses specifically on the creation of networked information resources in support of elements of the Synthesis program [Saylor, 1992], The Council on Library Resources is examining how to improve access to engineering information [Council on Library Resources, 1990; Council on Library Resources, 1992]; here, again, network-based access to information is viewed as playing a key role. Major scientific programs in areas ranging from Molecular Biology to Global Climate Change all view the development and operation of networked databases as essential program components [Olson, 1993; Stonebraker, 1992]. The higher education and research library communities have created the Coalition for Networked Information (CNI), a joint project of CAUSE, EDUCOM and the Association of Research Libraries (ARL) to specifically address

[1]This was signed into law by President Clinton on June 8, 1993 as Public Law 103-40

[2]While agencies such as NASA seem to have made a commitment to the **networked** information model of access to agency information, other groups—for example, the **Bureau of the Census—have addressed the distribution of government information through the publication** of CD-ROM databases, leaving it to the user communities (i.e. the universities or the Federal depository libraries) to determine how the information they publish on CD-ROM should be made generally accessible through the network. At present most CD-ROM databases are unsuitable for use in the networked information environment, despite efforts by group such as the Air Transport Association and the Intelligence Community Data Handling Committee to define network interfaces to CD-ROM databases [Bowers & Shapiro, 1992]. In still other cases federal agencies such as the SEC have formed alliances with the private section (Mead Data Corp., in the case of the SEC) to offer access to federal information through the networks [Love, 1993]. There is a great need for more consistent policies for access to government information through the networks.

[3]Utah makes legislative information available through the Internet.

[4]In California, Assembly Bill 1624, currently under consideration in the state legislature, would make legislative information available through network access.

[5]The Maryland State Librarian is leading a major effort to link libraries throughout the state using the Internet. Similar projects are under discussion in Virginia and other states, [Library of Congress Network Advisory Committee, 1992].

networked information issues [Peters, 1992a], as well as devoting substantial attention to broader networking initiatives within the programs of the parent organizations (for example, EDUCOM'S Networking and Telecommunications Task Force); the CNI initiative reaches out beyond the research and education community to reach providers of networks and information technology, publishers, and even, to a limited extend information users in the private sector.

There is a great deal of talk about "digital libraries", "electronic library collections", "electronic libraries", "electronic journals" and "network-accessible library collections"; such visions have captured the imagination of many scholars, educators, librarians, and policy-makers [Lynch, 1991a], and are increasingly attracting the interests of the commercial sector—particularly publishers, mass media, and other information providers—as potentially lucrative new markets. Indeed, the upcoming transition to electronic information resources is viewed hopefully by some as a means of relieving increasingly intolerable financial pressures on the entire system of scholarly publishing.[6] Yet the definition of a digital library remains controversial. Personally, I prefer to consider the viewpoint that stresses electronic library collections; a library is an organizaiion that acquires, structures, provides access to, and preserves information of all kinds, and within this context network-based electronic information is just another type of material that the library can acquire and manage (albeit one with many unique, intriguing and novel properties). Additionally, the availability of digital content and sophisticated information technology of course permit the library as an organization to offer new organizational and access services, and to move from a primarily "passive" organization to one that actively provides information to its users through interactive services and automated, network-based, content-driven information delivery.

When we speak of digital libraries, however, we invoke not only this concept of electronic library collections but also visions of the integration of networked information resources of all kinds (including, for example, numeric databases that libraries have typically neglected and remote sensing databases that are collected as part of various scientific programs outside of the library context) into new collaborative environments (co-laboratories) [Lederberg & Uncapher, 1989]; some term such collections of databases and related networked information resources to be digital libraries. There is discussion of coupling information technology with a new partnership among researchers, information management specialists and information technology experts

6 The primary source of these pressures is that libraries can afford to purchase a smaller and smaller part of the annual output of scholarly books and journals worldwide. The roots of this crisis are complex and form the subject of extensive debate between librarians, publishers, and academics. Many librarians argue that the prices for these materials are rising far in excess of the rate of inflation, in part due to price gouging by the publishers. The publishers submit that the size of the body of annual research is growing rapidly, and that library funding has not kept up with this rate of growth; they also identify other factors such as international currency fluctuations that have contributed to the problem in recent years. For a survey of some of these issues see [Grycz 1992]. A discussion of these issues is outside the scope of this paper; however, it is important to recognize that conversion of scholarly materials to electronic formats may reduce printing, distribution, handling and storage costs for these materials somewhat, but will generate new costs in retooling editorial and production processes and in investments in information technology and infrastructure. Overall, it is not clear that conversion to electronic formats will substantially reduce costs for scholarly materials, though it will undoubtedly redistribute these costs. Further, if, as some librarians argue, much of the problem is publisher profiteering, a shift to electronic scholarly publications will only alter the economics if it causes substantial changes in the system and the role of publishers—particularly commercial publishers—within that system.

under a model called “knowledge management” that relies heavily on networked information resources to directly support and integrate with the research process itself as well as the diffusion of new knowledge to the research community and the dissemination of research results [Lucier, 1990; Lucier, 1992]. I view these new networked information resources to be something fundamentally new, and different from library collections (though they might, in some cases, be part of a library’s collection, or a part of the services offered by a library); certainly they are different from libraries (as organizations), though in some cases libraries may be the organizations that create, manage or fund these new networked information resources. We will need a new terminology and taxonomy for the networked information age. But, in any event, the focus of this paper will be collections of network-accessible information resources and the roles of libraries in maintaining and providing access to them.

There are a vast number of experiments underway at present to use the network to deliver or provide access to bitmapped images of print publications, including document delivery services being offered by CARL, Engineering Index, University Microfilms, Faxon and others, often in complex business partnerships with traditional secondary (i.e. bibliographic) database access providers such as OCLC, RLG, or Dialog. Primary scientific journal publishers such as Elsevier and Springer-Verlag are conducting experiments with various universities under which they are making the contents of some of their journals available electronically to institutions either under site license or pay per view arrangements. In addition, various third-party aggregators and relicenses such as UMI and Information Access Corporation are now licensing full text or page images of journal contents for sizable collections of journals in specific areas directly to institutions, and a number of publishers are making the text of their newspapers, magazines, or other publications available for searching through database access providers such as Dialog or BRS on a transactional basis.

Indigenous network-based electronic journals are now well established, and their number continues to grow (though it is important to recognize that they are still a very minor force, compared to the existing print journal system, in most disciplines). The vast majority of these are made available for free on the Internet, and they include both journals structured in analogy to peer-reviewed print journals, such as Postmoderm *Culture* (which Oxford University Press has recently agreed to market in a parallel print version), *Psycology, Public Access Computer Systems Review* and many others (see Michael Strangelove’s bibliography [Strangelove, 1993] and other regular publications that are similar to edited newsletters *(Newsletter on* Setials *Pricing Issues, TidBits,* etc.). The edited newsletters are part of a continuum that fades off into “moderated discussion lists” implemented through LISTSERVERs or other mail reflectors, which might be viewed as continuously-published electronic newsletters that exploit the electronic medium of the network to avoid the need to gather submissions and commentary together into discrete issues. There are also thousands of unmoderated discussion groups which provide additional forums for discussion and information interchange; these have some elements in common with newsletters or other publications, but are really a new and unique artifact of the networked environment. Recently, a few for-fee journals have begun to publish either solely in electronic form (i.e. the OCLC/AAAS *Current Clinical* Trials experiment, a fully peer-reviewed journal [Keyhani, 1993; Palca, 1991], or the *Computist’s Communiquè* by Ken Laws, more of a newsletter), or in parallel print and electronic editions (e.g. John Quarterman’s Matrix

News). It seems probable that the development of for-fee journals on the network has been inhibited by publisher concerns about the acceptable use policies that govern traffic on much of the Internet; as it appears that the acceptable use policy may well either be abandoned or interpreted liberally enough to comfortably accommodate this type of network based publication and publishers can find other publishers distributing journals in the network environment without problems, the number of for-fee journals will grow rapidly.

Paralleling these initiatives in the creation of content, a great deal of effort is being devoted to the development of tools for network navigation and information retrieval. The development of standards for resource description, location, and access in a distributed environment are also a key part of the development of the tools themselves. Major efforts in this area include the Gopher project at the University of Minnesota [Alberti, Anklesaria, Linder, MaCahill, & Torrey, 1992; Wiggens, 1993], the World Wide Web system from CERN [Berners-Lee, Cailliau, Groff, & Pollermann, 1992], the WAIS system that was originally developed as joint project of Thinking Machines, Apple, Dow Jones and KPMG which is now being carried forward by a number of organizations, including the new NSF-funded Clearinghouse for Networked Information Discovery and Retrieval in North Carolina [Kahle, Morris, Davis, & Tiene, 1992a], the archie system developed at McGill University [Emtage & Deutsch, 1991; Simmonds, 1993], the resource discovery work carried out by Mike Schwartz and his colleagues at the University of Colorado, and others [Schwartz, Emtage, Kahle, & Neuman, 1992; Schwartz, 1989; Schwartz, Hardy, Heinzman, & Hirschowitz, 1991]. Recently, the National Science Foundation awarded a sizable contract to AT&T for the development of directories for the Internet; while this contract is primarily to compile and operate such a directory using existing technologies and standards, and the resource directory being developed does not seem to incorporate any sophisticated classification or retrieval approaches, the AT&T effect is likely to focus and stimulate further effort in this area. More research-oriented work is also underway in developing cataloging and directory tools through the CNI Topnode project,[7] the Department of Education funded project for cataloging Internet resources at OCLC [Dillon, 1993], the work of the Library of Congress on cataloging practices for electronic information resources [Library of Congress, 1991a; Library of Congress, 1991b; Library of Congress, 1993], and the efforts of various working groups within the Internet Engineering Task Force on document location and identification.[8] Other important standards work includes activities such as the Z39.50 Implementor's Group, which is addressing both functional extensions to the Z39.50 computer-to-computer information retrieval protocol, a basic tool for searching information resources on the network, and also attempting to resolve interoperability issues as the Z39.50 protocol moves toward widespread implementation [Lynch, 1991b; Lynch, Hinnebusch, Peters, & McCallum, 1990]. In addition, of course, standards developed within broader communities to describe various types of record and document interchange formats are of critical importance to the development of networked information retrieval tools.

[7] Information on the current status of this project can be obtained from the Coalition for Networked Information, or by searching CNI's file server at ftp.cni.org.

[8] While not much has been published on this yet, the IETF should be issuing a series of draft RFCS within the next six months. The general approach that is being pursued is described in Section 13 of this paper.

Institutionally based projects at universities such as Carnegie-Mellon (Project Mercury) [Arms, 1992], Cornell (various projects) [Lesk, 1991], the University of California (various projects) [Lynch, 1989; Lynch, 1992]. The University of Southern California Watkins, 1991] have focused on developing systems for the storage and delivery of electronic information resources to institutional user communities, in some cases integrating and building upon tools and standards developed on a national and international level. Some other projects, notably at Yale, Cornell, Barry Shein's Online Book Initiative (hosted at world. std.com) and Michael Hart's Project Gutenberg are working with public domain collections (either out of copyright materials or government material not subject to copyright) as prototypical electronic library collections. In some ways, these out of copyright collections are very attractive test cases as they permit the institution to focus on the technology and end user requirements of the application without becoming enmeshed in economic and legal (intellectual property) concerns.

Not all work on networked information access is rooted in the higher education and library communities or the efforts of commercial firms that primarily serve these communities. For example, a number of communities have deployed versions of the Freenet system developed by Tom Grunder in Cleveland, Ohio [Grunder, 1992; Watkins, 1992]. This is a system which provides the electronic analog of a town, complete with a town hall, libraries, schools, medical facilities, discussion groups, and other areas. While some implementations have been supported in part by libraries and/or universities, Freenets may equally well be sponsored by municipal governments or private citizen groups outside of the higher education and research communities. In addition, commercial services such as CompuServe and America Online are now well established and offer access to sizable collections of information; their primary user communities are outside of the academic or library worlds.

Recently, Carl Malamud established a project called Internet Talk Radio which offers a mixture of interviews and live coverage of speeches and other newsworthy events; the content is somewhat similar to that of the C-SPAN cable network, although it includes announcements from commercial underwriters similar to those found on Public Television (not really full scale advertising by sponsors) [Markoff, 1993]. Internet Talk Radio has coverage into the National Press Club in Washington, DC and is scheduled to "broadcast" its first coverage of a congressional hearing later this summer, Internet

[9] It should be noted that, while out of copyright material will be a very important resource for libraries that wish to explore the electronic storage and dissemination of material exactly because this material is not subject to copyright constraints, such material is substantially difficult to identify; worse, the identification of such material is growing more complex as the issues are explored in more depth. Consider first simple textual materials. In the US, currently, the period during which a work is subject to copyright is defined by the life of the author plus a certain number of years, rather than the old definition which was just a fixed number of years from the date of copyright. This means that a library that wants to determine whether a work is still under copyright protection must attempt to determine whether the work's author is still living or when he or she died. This is a major research problem. Further, international copyright issues have become extremely complex. For example, Project Gutenberg recently made a copy of Peter Pan available, since the work appeared to be out of copyright in the US, only to subsequently discover that there is apparently a special exemption for this work under UK copyright law that permanently assigns copyright protection to this work and donates the proceeds to a hospital in the UK, and thus the electronic text could not be distributed in the UK legally. How this strange exemption in UK copyright law relates to the Berne Convention and the internationally reciprocal copyright agreements that the US has agreed to honor is a subject for legal scholars that I will not speculate upon here; however, it is a good illustration of the problems of identifying material that is no longer subject to copyright.

Talk Radio captures audio from these events and distributes it over the Internet in real time using multicast technology to interested listeners; in addition, digital files containing the audio for the broadcast events are archived on the network and can be retrieved and played by individuals with appropriate audio boards in their machines at any time on demand. Even for simple audio, these files are quite large and stretch the capabilities of many machines on the net. The Internet Engineering Task Force has been experimenting with digital audio and video distribution of key sessions at its meetings using similar multicasting technology, though this is considerably more taxing for the network due to the data volumes involved. As the network capacity expands and the supporting capture, playback and multicasting technologies become more mature and more widely available it seems likely that this type of audio and video coverage of events of interest, both multicast real time and stored in archives for later demand playback, will become more commonplace.

Yet, despite this rich and wide-reaching series of projects (and what has been described here is only a sampling intended to give the reader a sense of current developments) which we hope will yield knowledge and insight that will inform future efforts,[10]" there is little consensus about the future structure of electronic libraries, digital libraries, network-accessible libraries or whatever one wants to call them-or even if these terms refer to the same things. Some people refer to collections of network-accessible files as a digital library; this is common in some parts of the computer science community, for example. Some from the publishing community speak of digital libraries when a perhaps more accurate term might be a digital bookstore or information store. Those viewing the evolution of electronic information resources from the library tradition tend to think of networked information as simply one more component of a traditional library's collection, subject to the basic library functions of selection, organization, provision of access, and preservation, suitably adapted for the unique characteristics of the network environment (for example, you can select a network resource that you provide access to without physically making it a part of a given library's collection—in other words, performing acquisitions without taking physical possession, as distinct to providing some form of subsidized access to a resource that the library continues to regard as "external" and available through mechanisms such as interlibray loan or short term contract to subsidize access in the networked environment). Indeed, with the network's ability to dissolve geographic boundaries and unify access to autonomous, physically and organizationally distinct resources, fundamental questions are being raised about the nature of these future electronic information collections—for example, might there just be one logical "library" of information for each discipline [Loken, 1990], perhaps managed by a professional society, in the sense that the user would communicate with only a single access-providing organization for the discipline's literature?

---

xtent to which current prototypes will in fact help to resolve the open questions is problematic. Many of the prototypes are being rushed into production use, without any systematic evaluation of the human or economic outcomes. Too often there is funding to build prototypes, but no funding to evaluate them rigorously. In some cases the economic viability of projects beyond the prototype state is unclear, and there is a real lack of common economic models that permit comparisons to be drawn between projects. The definitional difficulties surrounding the concept of "digital libraries" are indicative of the severity of this problem.

3. An Overview of Access and Integrity Issues in the Networked Information Environment

The institutions that are libraries—be they public libraries or research libraries—have addressed a number of concerns about the accessibility and integrity of printed information that arise from diverse quarters ranging from the needs of the academic community to manage and provide access to the scholarly record through the needs of the government to ensure the existence of an informed citizenry with access to vital government information resources. Libraries ensure a certain base level of access to information irrelevant of the financial status of the information seeker. Many of these concerns do not have well-established, clearly defined constituencies or clearly stated requirements. But the concerns are nonetheless real, and of vital importance to our nation and our society. As the nature of the information changes from printed pages to network-accessible information resources, we can no longer assume that old mechanisms will continue to work, or that they will be adequate to address the full range of new issues that are raised by electronic information resources; indeed, we do not yet fully understand the implications of a large scale shift to electronic information or the new roles that we will expect libraries to undertake in this context.

This paper examines a series of specific issues related to the access and integrity of electronic information in a networked environment, and current and potential roles that libraries and other institutions may play in addressing these issues. It also explores the ways in which the transition to the networked information environment may call existing library practices and roles into question.

Access to information in a networked environment is an area that is often treated very superficially. There is a tendency to incorrectly equate access to the network with access to information; part of this is a legacy from the early focus on communications infrastructure rather than network content. Another part is the fact that traditionally the vast bulk of information on the Internet has been publicly accessible if one could simply obtain access to the Internet itself, figure out how to use it, and figure out where to locate the information you wanted. As proprietary information becomes accessible on the Internet on a large scale, this will change drastically. In my view, access to the network will become commonplace over the next decade or so, much as access to the public switched telephone network is relatively ubiquitous today. But in the new "information age" information will not necessarily be readily accessible or affordable; indeed, if information is to become the new coin of the realm, there is no doubt in my mind that there will still be the rich and the impoverished-though even the impoverished may have a relatively high standard of access to information, compared to today's information poor in the US, or tomorrow's information poor globally .11 This

---

11 The information poor are not the same as the illiterate. The illiterate are a group that lack specific training and skills increasingly essential in modern society; the information poor may not necessarily lack the skills to find or comprehend the information they need, but rather may simply not be able to afford to pay for access to information. While illiteracy is a problem that is often the result *of* poverty, it is really the lack of a specific skill. Lack of access to information is a condition that is created by at least in part by poverty (and really more generally by a gap between the price of information access and the economic conditions of the person who needs to obtain access to information), but which is at least to some extent rectified by subsidizing information access, as opposed to illiteracy, which usually is not the result of inability of the illiterate to obtain access to printed material. Information poverty is a mix of two factors:

paper will focus on information access issues and largely omit issues related to base network access. Access here will also be viewed in the broad sense; not only considering who can have access and how much they must pay, but when they can have access, and who knows what they are accessing.

Integrity of electronic information is another problematic area. Many people have a bias that leads them to view electronic information as less stable than printed information-electronic information is subject to unannounced revision by insidious parties, corruption by viruses unleashed on the network or by individuals breaking into computer systems. In fact, the issues here are extremely complex, ranging from the balancing of the ability of the network to support access to the most current information against the need to maintain a trail of citeable versions linked to specific points in time, through questions of long term preservation of digital information. It is interesting to note in this connection that many of our expectations about networked electronic information are derived from our experience with, and expectations about, print information, and that in fact we regularly accept completely different rules for broadcast mass media information than we do for print; similarly, much of the legal framework for electronic information (with the exception of some very specific counterexamples, such as integrated circuit masks) also has its basis in practices related to print materials. (It is also worth noting that most libraries have tended to avoid becoming much involved with providing access to the contents of broadcast mass media). Other issues in this area include the problems of intellectual property, hidden bias of many different types, and loss of information diversity available to the public. Our expectations about the integrity of electronic information are unclear; in fact, these expectations vary depending on the use we wish to make of a given electronic information resource.

Integrity and access are interrelated in complex ways; in the evolving context of networked information, the relationship is far more complex than in the existing print-based environment. In the electronic environment the balance of relationships between the creators of information, the distributors and rights holders (publishers), the stewards (libraries) and the consumers of information seem to be changing radically. Within the print literature framework each of these stakeholders had well-established roles in ensuring integrity and providing access; with the shift in relationships and responsibilities, these roles will also change. Access to electronic information is of questionable value if the integrity of that information is seriously compromised; indeed, access to inaccurate information, or even deliberate misinformation, may be worse than no access at all, particularly for the naive user who is not inclined to question the information that the new electronic infrastructure is offering. Further, certain characteristics of the mechanics of accessing electronic information in the networked environment may lead to new means of comprising the integrity of that information.

---

inability to afford access to information and lack of skills to obtain, navigate, and evaluate information. One might argue, for example, that many scientists and engineers are in fact information illiterate, although they can certainly afford access to substantial bodies of information; they lack the skills to utilize this body of information. Further, literacy, in an world that is increasingly full of multimedia information, may not be always be a prerequisite to being able to understand information once one obtains access to it. Indeed, in the future, the relationships between literacy, having the skills necessary to locate needed information and/or access to trained intermediaries such as librarians who can help to locate information, and having the ability to afford access to information (either by paying for it directly or thorough organizations like libraries that have historically subsidized access to information) are going to become much more complex, and deserve new attention in the context of the coming age of electronic information.

Conversely, even if the integrity of the scholarly or cultural record is guaranteed, such a guarantee is of limited value without a corresponding guarantee of continued access to such a record.

In discussing issues of access and integrity in networked information today, there is a very strong bias towards issues specific to scholarly information; this is to be expected, given that the academic and research communities have up until now been the primary constituencies on the Internet. These are relatively sophisticated communities, and communities with an ethos that is strongly oriented towards citation, attribution, and preservation of the scholarly record. Indeed, as one reviewer noted, this ethos ties these communities to the system of print publication, and emphasizes that networked information must offer the same integrity and access if it is to become an acceptable replacement for print. Scholars must be certain that their work IS accessible to their peers and that the integrity of their works is maintained. Yet if one examines the current growth of the Internet, the fastest growing sector is commercial organizations. Primary and secondary education and public libraries are one of the major potential growth sectors (if funding strategies can be found to pay for their connection and for the associated information technology and staff training investments that will have to be made within the institutions themselves). There is now discussion about the role of the Internet as a precursor to a National Information Infrastructure (which really might be more appropriately called a National Networking Infrastructure) which would connect an even broader constituency. As the networked community expands, we will see a continued shift in expectations and values about information access, integrity and management, and the appearance of new types of information that have much more in common with the traditional contents of print and electronic mass media today than the bulk of the information that populates the current Internet. This paper thus attempts to take a broader view of the access and integrity issues, and to view them in terms of the expanding constituencies and types of information on the network.

To help to further clarify the scope and focus of this paper, let me emphasize that the paper devotes very little attention to the important and currently vigorously debated questions about government information, and in particular what government information should be available to the public, under what terms, and at what costs. This is a public policy issue of considerable complexity in its own right, and has a number of specific implications for libraries, particularly in their roles as depositories of and access points to government information. If anything, the emphasis here is more on information created and owned by other institutions, such as publishers and the academic community. An aggressive government program which expands the base of publicly owned information that is then made available to the public widely, at little or no additional cost, could well begin to alter some of the trends and evolving balances that are discussed throughout this paper. These issues merit considerably more exploration and discussion. Hopefully, however, this paper will provide a basis for such discussion, since, to a great extent, issues of access and integrity of collections of information in digital formats, and the roles of libraries in organizing, preserving and providing access to these collections of information are independent of the information's source.

4. Access to Electronic Information: the Changing Legal Framework

From Copyright and Sale to Contract Law

In order to understand the massive shift that is taking place as libraries begin to acquire electronic information to supplement and perhaps ultimately replace much of the printed information they have traditionally purchased, it is first necessary to review the evolution of the system through which libraries support access to printed material to provide a basis for comparison.

To a great extent, the cornerstone for the success of libraries as institutions in the United States has relied on the fact that printed information is virtually always purchased by libraries, and thus its subsequent use is controlled by the copyright laws of the US and the doctrine of first sale. Without going into great depth (and the reader should note that the author of this paper is not an attorney, and is not attempting to give legal advice, but only his own understanding of the way that the involved institutions have typically interpreted the relevant laws), what this legal framework for print publications means is that a library, having purchased a book or journal is free to loan it to patrons or other libraries, or to re-sell the material.[12] Moreover, the use of this copyrighted material by the libray or its patrons is governed by the copyright law, which represents a public policy established by the Congress through legislation which is based on the constitutional responsibility delegated to the Congress to promote the useful arts and sciences. The promotion of these useful arts and sciences has historically required the Congress to achieve a balance between compensating the creators of intellectual property and recognizing the needs and interests of the public to have access to such intellectual property as a basis for further progress in these arts and sciences. (Note that in this section I will use the term “copyright law” to refer rather broadly to the actual legislation itself, and also the body of legal interpretation and legislative history and intent that surrounds and supports the actual legislation.). Under copyright law, while a library, for example, is free to loan the physical artifact it has purchased, it is restricted, in most cases, from duplicating the material. Certain provisions of the current copyright law such as the Fair Use provisions explicitly recognize, for example, the importance of permitting a scholar to quote from a copyrighted work (within limits) for the purpose of criticism or further scholarly analysis; the copyright law also recognizes the importance of permitting the use of copyrighted material in certain educational settings. The importance to society of preserving the historical record is recognized in the permission that is granted to libraries to make a single copy of a physically deteriorating out of print but still copyrighted work for preservation purposes if no reasonable substitute is available. Patent law, to cite another example of the balance between encouraging creators of intellectual property and the needs of the public to be able to make use of that property once it is created, can be viewed as a compact that trades protection of an inventor’s intellectual property

12 There is no international consensus on the details of the doctrine of first Sale; In some European countries, for example, I believe that the right to loan purchased materials can in fact be excluded from the set of rights passed to a purchaser (such as the right to resell the copy of the work that he or she has purchased) as part of the initial sale.

(during a limited period of time) for the disclosure of that property to the public as a means of moving the state of the art forward.

It is important to note that there is nothing, as far as I know, in the current copyright law that requires printed material protected by copyright to be sold to purchasers; there is no a priori reason why a contract could not be written by a rights holder which allows an organization to license copyrighted printed material for a limited period of time and for specific uses outside of the framework of the doctrine of first sale, and such a license might grant the licensee much more limited rights than the licensee would have had under copyright law if the material had been actually sold. In fact, certain print publishers (for example, publishers of some types of atlases, directories and other compilations such as Dun and Bradstreet) have attempted to use licensing as a means of controlling their products for years, and recently the library community has encountered a number of other cases where print publishers have attempted to restrict the use of their publications through license agreements.[13] In many cases, the validity and enforceability of these terms imposed by the publishers have been of ambiguous as the publishers seem to be following the "shrink-wrap" license model of the software industry (that is, a buyer "purchases" something and by opening the package, or at least by using the "purchased" product the publisher states that the buyer is implicitly entering into a legal agreement to adhere to the license terms stated in the license agreement enclosed with the product). In most cases the library community has resisted (and occasionally prominently ignored license terms) products that come with such license constraints. I am unaware of significant test cases that have come to litigation to clarify these situations; further, I would speculate that in many cases where publishers have a strong motivation to attempt to protect print publications with license terms that limit, for example, who can see the printed material, are for very high priced, specialized, highly time sensitive information which is more typically purchased by commercial corporations rather than the general library community (for example, information related to the financial or securities industries).

An interesting related issue is whether publishers are required to make available their publications to libraries,[14] and if so under what terms and constraints. It is a well established practice today for publishers to use differential pricing for libraries and individuals, often with a very wide price spread: a scholarly journal might cost a library $1000/year for a subscription, but the publisher may also offer individuals at a subscribing institution (such as a university) the option of purchasing supplemental individual subscriptions for only $150/year.[15] Another common variation is a journal

---

13 The a~$_{emp}$t t. contro[access to print material through license agreement is not new, although it is relatively unusual in the United States. See, for example, Library Journal Volume 1 Number 2 (1877) in which the practices of publishers are deplored.

14 From a strictly legal perspective, I understand that publishers are under no compulsion to offer their works to any specific individual or institution.

15 While Commercial publishers often offer discounted subscriptions to ***individuals*** as a benefit for belonging to institutions where a library subscription to a journal is held, some professional societies go even farther and offer discounted subscriptions not only to individuals but to academic *departments* (for example, for departmental libraries, which are typically funded at the department level) for additional copies of journals. Today, subscription costs for scholarly journals have escalated to the level where individuals seldom subscribe even at the discounted prices, but the departmental level offerings are very attractive to some academic departments. The American Mathematical Society, for example, has used a

that comes to each individual that is a member of a professional society as a benefit of membership, but that is only available for library subscriptions (since membership in the organization is on an individual, rather than institutional basis, or institutional membership is many times as expensive as individual membership).[16] In response to such pricing schemes a librarian at an institution is often "encouraged" to join the professional society to get the journal at a reduced rate to the institution, with the membership fee reimbursed to the librarian. The publishers and professional societies suggest that it is at the least unethical for an individual to act as a "front" for an institution by ordering an individual subscription that is really going to be placed in the institution's library for general use,[17] and sometimes ask individuals ordering under the individual pricing scheme to sign statements asserting that they will not place the copies they are ordering in a library (indeed, some publishers have gone so far as to affix stickers indicating "individual use only" to publications shipped under individual subscription rates), but the legal enforceability of this is unclear given that once the individual obtains the material it is subject to the doctrine of first sale. Yet another interesting variation on this theme **arises** with publishers of very costly, time sensitive material-for example, a market research report at $5000 a copy. Such a publisher might well choose not to market to libraries, and perhaps not even to sell to a library that placed an order (or to process such an order very slowly). It is unclear whether such a publisher could actually be compelled to make a sale to a library.[18]

Licensing and the Interlibrary Loan System in the United States

The doctrine of first sale has also had another critically important implication for the library community in the United States; it has allowed the development of the interlibrary loan (ILL) system which permits one library to borrow materials from another. Historically, interlibrary loan was implemented by the physical shipment of materials from one library to another, which is clearly permitted under the doctrine of

---

central library subscription to subsidize departmental copies of some of their publications for a long time. It should also be noted that differential pricing is not clearly a completely evil thing; in the case of the American Mathematical Society, for example, one can view the library as subsidizing increased access to the Society's publications (in the pre-computer networking era) by allowing the Mathematics department to obtain a second, easily accessible copy of the material for a very small co-payment, Of course, in an age of pervasive networks, where institutional copies of material should be as readily accessible to the institutional community as departmental or personal copies, the entire concept of differential pricing begins to break down quickly.

[16] It should be emphasized that what might appear to be "discriminatory" pricing that Costs libraries much more than individual is really much more complex than it appears on the surface. One view of this situation is that libraries are partially subsidizing individual or departmental access to the literature for the cost of a small "co-payment". Another point to be considered is that publishers are making material available to individuals for the marginal costs of additional copies in print runs, discounted for the advertising that they can sell because of the individual readers of the journal. Differential pricing for libraries actually offers some benefits to libraries, institutions, and individual subscribers at those institutions.

[17] Arguably this could be regarded as fraud; while the argument that it Is unethical **IS Clear, it** is **unclear** whether it is' really illegal.

18 Ironically, many of these publishers of very expensive reports file copies of their material with the Library of Congress for copyright reasons, and this material is often publicly available there (though not always on a very timely basis).

first sale (though recently technological innovations, ranging from inexpensive xerography to facsimile transmission and more recently image transfer across the Internet have greatly complicated the picture and moved libraries into areas that are of much less clear legality). Libraries—and particularly research libraries—in the US are linked by an elaborate, complex set of multilateral interlibrary loan agreements. In many cases libraries have traditionally simply agreed to reciprocal sharing without charge to either library; in other cases (which are becoming more frequent as the size of the total body of material grows and the ability of individual libraries to locally acquire even a significant portion of this body of published material diminishes, leading to a massive explosion in interlibrary loan) the supplying library charges the borrowing library a fee, which may or may not be passed on to the patron at the borrowing library that initiated the interlibrary loan request. (Currently, even many public libraries assess patrons a nuisance charge—perhaps a dollar per book—for interlibrary loan, more as a means of controlling frivolous requests than anything else.) In cases where one library recharged another, there were often resource sharing funds established through agencies such as state libraries which helped to subsidize the costs of interlibrary loan (either by directly compensating the lending library for its costs in servicing interlibrary loan requests, or by compensating borrowing libraries for interlibrary loan charges that they incurred), at least in the case of public libraries as borrowers or lenders, leading to patterns where research libraries (for example, at universities) acted as the providers of last resort to public libraries within a state or even nationally.

To provide some quantitative sense of the size, cost and importance of the interlibrary loan system in the United States, consider the following figures from a recent Association of Research Libraries study on interlibrary loan [Baker & Jackson, 1992; Roche, 1993]. ARL's figures indicate that among the 119 ARL libraries in the US and Canada, interlibraryborrowing has increased I08% between 1981 and 1992; lending has grown 52% in the same time period. Recently, the growth rate in these areas has accelerated: lending has grown 45% from 1985-6 to 1991-2, and borrowing 47% from 1985-6 to 1991-2. As indicated, much of the driving force for this growth has been the increasing inability of libraries budgets to acquire published materials: since 1981, ARL library materials budgets have increased 224% while collections grew only by 12%; during this period ARL tracks the average cost of a book as rising 49%, and the cost of a journal subscription I09%. The average cost among ARL libraries for an ILL lending transaction is now about $11; for an ILL borrowing transaction, the cost is about $19 (note that these are costs, and not necessarily what one library charges another for such a transaction). Current estimates suggest that the US library community spends around $70 million per year performing ILL transactions.

The interlibrary loan system essentially allows libraries, as a community, to purchase expensive and/or lightly used printed materials and share them within the community, though of course the specific library that has actually purchased the material can offer better access to it than other libraries which may have to obtain it on behalf of their patrons through interlibrary loan. Interlibrary loan has historically been a rather slow, expensive, and inefficient process. As use of ILL has increased due to the diminishing ability of any given library, even a world-class research library, to hold the majority of the published literature, considerable attention has been focused on speeding up and

improving the cost efficiency of the interlibrary loan process.[19] Work in this area has ranged from the use of electronic ILL systems linked to large national databases of holdings (such as OCLC) which allow a requesting library to quickly identify other libraries that probably hold material and dispatch loan requests to them through the exploitation of technology to reduce the cost of the actual shipment of material. The first step in this latter area was for the lending institution to send a Xerox of a journal article rather than the actual journal copy, so that the borrowing library did not have to return the material and the lending library did not lose use of it while it was out on interlibrary loan. This use of photocopying technology was controversial, and libraries ultimately agreed to limit it through the CONTU guidelines, which define limits on the number of articles that a borrowing library can request (from whatever source) from a given journal per year without additional payment to the rights holder (either through article-based fees assessed through organizations such as the Copyright Clearance Center or by subscribing to the journal directly); while the CONTU guidelines are generally accepted practice, they do not have any legal standing that I am aware of, and have not been subjected to any test cases—they merely represent an ad hoc, informal agreement between the library and publisher communities as to the bounds of acceptable behavior by libraries. More recently, libraries have employed both fax and Internet-based transmission systems such as the RLG Ariel product to further speed up the transfer of copies of material from one library to another in the ILL context, and with each additional application of technology the publishers have become more uncomfortable, and more resistant (with some legal grounds for doing so, though again this has not been subject to test). Interestingly, over the past two years we have seen the deployment of a number of commercial document delivery services (the fees for which cover not only the delivery of the document to the requesting library but also copyright fees to the publisher) offering rates that are competitive—indeed, perhaps substantially better-than the costs that a borrowing library would incur for obtaining material such as journal articles through traditional interlibrary loan processe.[20] At the same time, the research library community is mounting a major effort under the

19 It is important to recognize that in the United States the interlibrary loan system is a distributed sYstem in most cases; that is, individual libraries make their own decisions about which libraries they will use as potential interlibrary loan providers. Libraries set their own rates for interlibrary loans typically, though in some cases there are consortium arrangements that provide fixed rates for libraries that are members of the consortium. This is in contrast to the situation in many other countries, where there is a national library that acts both as the coordinator of the country's ILL system and usually as the lending library of last resort for the national ILL system. In the US, such arrangements only occur in the biomedical and health sciences, where the National Library of Medicine is designated to act as a national library, and to a lesser extent in the agricultural disciplines, where the National Agricultural Library often acts in this role, although its primary mandate is to function as a library supporting the US Department of Agriculture rather than the agricultural community as a whole. The Library of Congress is not a national library and does not serve this function for interlibrary loan. In some states, such as California, the state research university (the University of California, the case of California) serves as the library of last resort for borrowing libraries within the state, though often coordination of in-state interlibrary borrowing patterns is done by the State Library rather than the libraries of last resort for interlibrary loan.

20 The economics of interlibrary loan are actually quite complex, in the sense that lenders of material through interlibrary loan frequently do not recover their full costs for providing material, and, in fact, it is a matter of considerable debate whether they can even identify what these costs are; although the ARL study provides an important step in this direction, it focuses on one specific type of library. One can look simply at how a library trying to obtain material for a patron can do this most cheaply, but perhaps a better perspective would be to consider the overall cost to the library community as a whole in supporting interlibrary loan as opposed to using commercial document delivery services.

auspices of the Association for Research Libraries to improve interlibrary access to materials. It remains to be seen to what extent the new commercial document delivery services supplant traditional ILL in the 1990s.

Now, consider a library acquiring information in an electronic format. Such information is almost never, today, so/cl to a library (under the doctrine of first sale); rather, it is licensed to the library that acquires it, with the terms under which the acquiring library can utilize the information defined by a contract typically far more restrictive than copyright law. The licensing contract typically includes statements that define the user community permitted to utilize the electronic information as well as terms that define the specific uses that this user community may make of the licensed electronic information. These terms typically do not reflect any consideration of public policy decisions such as fair use, and in fact the licensing organization may well be liable for what its patrons do with the licensed information. Of equal importance, the contracts typically do not recognize activities such as interlibrary loan, and prohibit the library licensing the information from making it available outside of that library's immediate user community. This destroys the current cost-sharing structure that has been put in place among libraries through the existing interlibrary loan system, and makes each library (or, perhaps, the patrons of that library) responsible for the acquisitions cost of any material that is to be supplied to those patrons in electronic form.[21] The implications of this shift from copyright law and the doctrine of first sale to contract law (and very restrictive contract terms) is potentially devastating to the library community and to the ability of library patrons to obtain access to electronic information—in particular, it dissolves the historical linkage by which public libraries can provide access to information that is primarily held by research libraries to individuals desiring access to this information. There is also a great irony in the move to licensing in the context of computer communications networks-while these networks promise to largely eliminate the accidents of geography as an organizing principle for inter-institutional cooperation and to usher in a new era of cooperation among geographically dispersed organizations, the shift to licensing essentially means that each library contracting with a publisher or other information provider becomes as isolated, insular organization that cannot share its resources with any other organization on the network.

Other Implications of the Contract Law Framework for Libraries

The shift from copyright law to license agreements has a number of other implications, all of them troublesome. At a public policy level, the ability of the Congress to manage the balance between the creators of intellectual property and the public has been undermined since copyright law no longer defines this balance; rather it is defined by specific contracts between rights holders and libraries. From the legal perspective, there is a very complex and ambiguous area having to do with the preemption of contract law (defined, at least in part, at the State level) of provisions defined by Federal (contract) law. At the level of the individual library writing contracts for

21 It is important to recognize that there are several conflicting and perhaps equally legitimate viewpoints here. Some publishers are viewing the transition to contract law as an opportunity to address what they view as an interlibrary loan system that has been pushed to the limit, well beyond where they are comfortable: they view some of the fax-based interlibrary loan resource sharing arrangements in force today as going well beyond the legislative mandate for shared access to information.

information in electronic form, the implications are even worse. Very few contracts with publishers today are perpetual licenses; rather, they are licenses for a fixed period of time, with terms subject to renegotiation when that time period expires.[22] Libraries typically have no controls on price increase when the license is renewed; thus, rather than considering a traditional collection development decision about whether to renew a given subscription in light of recent price increases, they face the decision as to whether to lose all existing material that is part of the subscription as well as future material if they choose not to commit funds to cover the publisher's price increase at renewal time. (In this context, it is important to recognize that price increases of 50°/0 or more at renewal time for electronic information are not uncommon, and that publishers are offering libraries various types of one-year free trials or other special introductory offers; given that all evidence is that electronic information, from a patron perspective, is among the most attractive and heavily used offerings of many libraries, and that in a large number of areas a given provider has what is essentially a monopoly position with no effective competition, no substitutable alternative available to the library-leaving the library with little choice but to seek the funds to pay for very large, unexpected, and, from the library's perspective sometimes extortionate cost increases). In a bad budget year a library might cut back on its purchases and subscriptions in the print environment, relying on its existing collection and ILL for new material that it cannot afford to purchase in the bad year; in the new electronic environment, a bad budget year might well cause the disappearance of much of the existing collection as well as affecting patron access to newly-published information.

The most common licensing situation today is for information that is either stored on some physical medium (such **as** a CD-ROM) which is then housed in the library or information provided on tape which is then mounted on computer systems belonging to the library. But, in fact, various types of usage restrictions defined by license agreements apply equally to remote databases where the library has contracted for access through the network. Indeed, some of the traditional database producers who offer access to their databases through online services like Dialog have a long-standing tradition of introducing complex and odious contractual terms[23] that predate any

---

22 One should not make too much of the problem of limited term as opposed to perpetual licenses. Some electronic information vendors are already incorporating perpetual license terms in their offerings, and several of the major publishers of primary journals have indicated a willingness to at least discuss a permanent licensing framework that parallels print practice. But this is an issue that libraries need to be aware of, and which does definitely have budgetary implications.

23 T. provide a sense of the contractual restrictions that one might encounter, consider the following terms excerpted from a recent set of contracts drafted by Dun and Bradstreet Inc. (D&B) and distributed to members of the Association of Independent Information Professionals, an association that represents free-lance information searchers that often work under contract to large companies. These terms are part of a rather complex legal agreement involving an independent information professional (IIP), and an end user employing the IIP, as well as D&B itself, and govern use of D&B information retrieved from Dialog. The terms include the following provisions:

- The information may only be provided by the IIP to a single end user.
- The II P may not overlay additional data elements to information retrieved from the D&B files, or merge this information with information from other sources (thus adding value). The information cannot be used to create a whole or any portion of a mailing or telemarketing list, or any other marketing or research aid or data compilation sold by the IIP to a third party.
- Only certain specific D&B files mounted on Dialog can be searched.
- I I P may not provide information to any person the II P has reason to believe will use it to engage in unfair or deceptive practices, or that will republish, subsequently license or resell the information.

significant use of these databases by research or public (as opposed to corporate) libraries.[24]

Practically speaking, no library is going to negotiate thousands of contracts, and no publisher wants to maintain contracts with thousands of libraries. This means that an industry of rights brokers will come into being. These may be aggregators such as University Microfilms (UMI) or Information Access Company; they may be clearinghouses similar to the Copyright Clearing Center (CCC) in the print world. Utilities may come into being that provide aggregated access to material provided by publishers, just as companies such as Dialog provided access to databases in the 1970s and 1980s; OCLC, among others, is already positioning for such a role through its mounting of various electronic journals such as *Current Clinical Trials.*

The uncertainty and restrictions surrounding contracts for access to electronic information are not the only problems that libraries will face in this transition from copyright law and purchase to contract law. Today, at least, there is no standardization of contracts, and efforts such as the CNI READI project [Peters, 1992b] that have sought to explore the potential of such standardization have been discouraging. Given that a large research library may well deal with several thousands of publishers in a given year, one can quickly see that in the electronic information environment, such a research library will be in no position to negotiate several thousand unique contracts with publishers for electronic information resources. Further, it is not just setting the contracts in place, which, as discussed, can be addressed to some extent through rights brokerage organizations (though this may impose a new cost on the overall system that is largely absent in print publication). Imagine the plight of a library that is attempting to support its host university's decision to enter into a cooperative agreement with a small not-for-profit research center which includes access to the university's library: there are now potentially thousands of separate contracts that need to be reviewed before the library can understand access constraints on this new cooperative venture; even in the case of rights brokerage organizations, unless there is a high degree of uniformity in the set of rights that the broker is prepared to license from various rightsholders and also in the terms of these licenses. In the most absurd case, a request to have access to a specific electronic information resource at the university library might well become a matter for the university's general counsel to

- The IIP acknowledges that D&B may introduce identifiable erroneous names that permit D&B to audit information use compliance, and the IIP agrees not to remove any such erroneous names.
- The end user agrees not to use the information as a factor in establishing an individual's eligibility for credit or insurance to be used primarily for personal, family or household purposes, or for employment.

24 One particularly sore point has been prohibitions on independent researchers who act as contractors for companies. The Association of Independent Information Professionals (AIIP) has spent a great deal of time over the past few years attempting to negotiate permissions to allow its members to legally use these databases on behalf of their clients. AIIP members, as independent small business people, tend to be meticulous about the legality of their database usage; the complexities that such usage restrictions would create for members of the academic community are so profound that in the academic environment it is likely that the restrictions on use would either be largely ignored, or the restricted use databases simply wouldn't be used. In an competitive environment this would be to some extent self-correcting; the restrictions would reduce revenue, and the competing product with the most liberal usage restrictions would be likely to gain market share. But, in an environment where the business community is the predominant revenue producer and the database is a relatively unique resource that is also needed by the research community, such restrictions quickly begin to function as a major barrier to access.

consider, and might take weeks or even months to resolve. This is a great distance from the old world in which a library could make its own decisions about who had access to its collection, and could readjust these policies in response to changes in the agreements that its host institution entered into.

The ability of publishers to specify usage conditions in a contract can also create other complex liabilities and administrative problems for libraries. For example, a publisher can specify that licensed material is being provided only for research and teaching applications. This can create legal ambiguities (and thus potential liabilities) if, for example, a faculty member uses some of this material in a book that is being commercially published, even if the use of this material might normally have been covered under the fair use exemptions for criticism within the copyright framework. Other information providers restrict printing of their material, downloading into personal databases, or the sharing of information with third parties; many of these restrictions are hard to define precisely, impossible for the library to enforce, and probably unrealistic. Given that the complexity of the usage restrictions is defined only by the publisher's ingenuity, we should not overlook the possibility of very odious usage restrictions being incorporated in contracts, or the liability that such contracts may create for libraries.

In providing a balanced view of the transition from copyright law to contract law it is important to recognize that institutions acquiring intellectual property are not simply being bullied into accepting contracts. While librarians may feel some discomfort about the transition, senior executives in most large organizations (for profit corporations or educational institutions) are relatively conservative and risk-adverse. The definition of rights under copyright law is interpreted through court decisions (case law) and thus appears to these decision-makers as fraught with ambiguity and uncertainty. Events such as the recent Texaco decision highlight the risks that institutions operating under copyright law can face. Contract law is less ambiguous and less risky as a framework for acquiring intellectual property, assuming that one has the money to pay the rights holders; this is a real attraction to management at many institutions.

The shift from copyright law to contract law seems inexorable, barring major restructuring of the copyright law. While consideration of such restructuring is outside the scope of this paper, which is focused on extrapolating and illuminating the implications of current trends within the current legal framework, it should be noted that a number of proposals have been developed for alternative legal frameworks. One example is compulsory licensing of materials in electronic formats. Depending on how fees were set in such an environment, some of the problems with current trends could be addressed, although such a shift would give rise to a number of new problems and issues, both at the public policy and practical operational levels.

Libraries and Multimedia: New Legal Issues

An increasing amount of electronic information will be in various multimedia formats. Here, a new and complex problem arises that is already causing great concern in **the** library community; the initial source of that concern is videotapes, videodisks and related materials. When dealing with printed material, the typical concern is over the right to make copies for various purposes. Loaning of physical artifacts is not much of a

question. Certainly, the ability of a library patron to view a book owned by a library is not an issue. But, if one studies the copyright law, one quickly finds that it addresses not only the right to copy, but also rights to display and perform works. While these are not much of an issue in respect to printed materials circulated by libraries, they represent real and complex problems for multimedia works. Without reviewing the legal details (which appear to be highly controversial at the present time) it does seem clear that the role of libraries in providing access to multimedia works is at best ambiguous. Even when operating under the general framework of copyright, as opposed to contract law, the worst case may be that libraries can acquire material that they cannot legally provide their patrons with the facilities to view within the library. (For a good summary of these issues from a legal perspective, see [Cochran].) Given the growing complexity of the technology base need to view (and interact with) certain types of multimedia products, this represents a very real barrier to access.

The issues related to traditional audiovisual materials are already serious, and have been a source of major problems for libraries. Early experiences with the lending of software has also revealed numerous issues. But perhaps even more important is the unresolved extent to which rights of performance and display will be attributed to the viewing of electronic information of all types, ranging from the browsing of bitmapped images of print pages through interaction with a digital movie driven by a program.

The widespread development of multimedia authoring tools will raise other issues as well, perhaps less for libraries than for users of digital information on the network. Multimedia integrates film clips, images, music and sound along with other content, and most developers of multimedia are not simultaneously artists, composers and musical performers. There will be a great demand for copyright-free (public domain) materials that can be included in multimedia works.[25] Here again one encounters a large number of ambiguous questions related to copyright law. One can find numerous opinions on these questions, but only very limited consensus and even less certainty. The questions include:

- Who owns the rights to an image? This includes photographs, images of classic paintings, and other materials? This is a particularly vexing question in regards to paintings owned by museums, for example. It's important to recognize that this is not a new problem that has been created by the digital environment; ownership of images is a very complex (and sometimes ambiguous) area in copyright law even in the older print and photographic environments. Digital imaging technologies and network distribution simply underscore existing uncertainties.

- If an image is digitized, and then perhaps subsequently enhanced, is this protected under copyright?

---

25 We are already seeing the beginning of this process in the very complex restrictions that accompany demo disks for products such as Apple's Quicktime and the various "clip art" and "stock photo" products being offered to multimedia developers. People want to incorporate bits of this material into all sorts of new multimedia: sales presentations, educational materials, demonstrations, and training materials. In many cases it is unclear when rights must be cleared. It seems likely that a few well-publicized legal actions could lead to an atmosphere of pervasive paranoia that might quickly retard the use of multimedia technologies, particularly by the business and educational communities.

- To what extent is the linkage of a series of media (for example, images and a sound track) copyright able separately from the images themselves and the sound track itself?

- If an out of copyright text is scanned or keyboarded and then edited, to what extent is this protectable under copyright?

- How does the developer of a multimedia product, attempting to comply with the law and to behave responsibly, determine whether component works that he or she wishes to incorporate into a multimedia product are protected by copyright?

- To what extent are libraries (or other networked information providers) liable for contributing to copyright infringement in an electronic information environment. To give only one example, a number of libraries are currently considering how to upgrade their facilities to accommodate users of portable notebook computers in conjunction with the overall move to electronic information. If a library permits patrons to connect to a library network to download images from the library's collection, to what extent is the library liable if these images prove to be copyrighted?

The problem here is again in large part the uncertainty. As matters stand today, many of these questions will have to be decided, at least as I understand it, through test cases in court. Most libraries (and their parent organizations, such as universities) are not eager to serve as such test **cases.** It is quite possible that attempts by libraries to limit the potential legal liabilities of the current uncertain copyright framework may also contribute to the destruction of the interlibrary loan system through a migration to acquiring material under license (contract law); understandably, most organizations will place a greater priority on managing their own legal exposures than they will on the ability to share their material with other organizations. Of course, there are alternatives to defining the specifics of copyright in the electronic environment through case law: these include both specific legislative actions that clarify and perhaps further define the law in the electronic information world, or joint agreements between information providers and information purchasers similar to the CONTU efforts which establish community guidelines without having the actual force of law.

As one reviewer of an draft of this paper reminded me, it is also important to consider the perspective of the individual developer (perhaps in a very modest, non commercial sense) or user of multimedia—the teacher in the classroom, the marketing representative preparing a sales presentation, or the individual citizen amusing him or herself. As has been illustrated by such issues as the transcription of phonograph records or more recently CD-Audio recordings to audio tape, or the taping of shows from the broadcast media for later viewing, many such users are literally unwilling to recognize or worry about legal restrictions to actions that seem to them to be reasonable. Such individuals—the majority of the citizens of the US—are likely to become outlaws (most often unwittingly, but occasionally as a matter of deliberate choice) outside of the framework of institutions and institutional liability rather than abide by complex, hard to understand legal restrictions that seem intuitively senseless. Laws that are strongly at odds with social norms are frequently bad laws, as they undermine people's acceptance of the overall law of the Iand. In areas such as personal taping (so that one can play the Audio CD one just purchased on one's car stereo, or watch a program that was broadcast again from one's VCR) my sense is that

the courts have largely upheld community consensus about what is reasonable, and sided with the consumer; any other choice would be impractical, to say the least. We will face many of the same issues in digital multimedia.

5. The Role of Secondary Information Sources and Automatic Indexing in Access to Electronic Information

Today, universities and other organizations are licensing and mounting abstracting and indexing (A&1) databases as adjuncts to their online catalogs; these databases provide library patrons with logical access to the journal literature (and sometimes also book chapters, books, technical reports, and other material) in a given discipline. A&1 databases contain citation information—authors, titles, journals of publication, page numbers—for material such as journal articles; in addition, they often include subject headings or other access terms assigned by indexers, and sometimes also abstracts of the contents of the articles, book chapters or other materials. Abstracting and indexing database records fill a role similar to that of cataloging records for books in a library catalog, but often provide more information about the work than a library catalog will. In general, library catalogs focus on monographic material-books, maps, sound recordings, films-or entire periodicals (for example, recording the fact that the library had a subscription to a given journal); abstracting and indexing databases typically focus on articles in journals or chapters in books.

Currently one of the major challenges for libraries is bridging the gap between the intellectual access offered by abstracting and indexing databases and access to their physical journal collections (as described in their catalogs thorough records of which journals they hold); in future the abstracting and indexing database providers will also offer links to electronic publications directly. The compilers of these databases wield great power that is just now being fully recognized. The experience of libraries in mounting online catalog databases (which typically cover only the monographic literature held by a given library) has been that when only the online catalog database was available some patrons tended to use monographic material almost exclusively; other (arguably more sophisticated) patrons who recognized that the journal literature was vital to their discipline tended to reject the online catalog as irrelevant. Indeed, this reaction to online catalogs was one of the primary forces that motivated libraries to license abstracting and indexing databases to attempt to bring access to the journal literature into balance with the access that they already offered to the monographic literature. Now that A&I databases in various disciplines are readily available to library patrons[26] these effectively define the relevant literature in these disciplines both in their

26 A few Points should be made about the origins and development of abstracting and indexing databases, and the impact of their conversion to electronic formats. In the mid 1800s various individuals and organizations began to compile indexes to parts of the journal literature and market these to libraries; however, the size of the journal literature was sufficiently small until the early 20th century so that at least large research libraries could actually create article-level card catalog entries for articles in journals to which they subscribed. Thus, up until the early 20th century, the library catalog served as a record of material that the library held, and specialized indices served as a means of providing access to the entire published literature in an area (whether the library owning the index owned the material or not). With the explosion of publication during the later part of the 20th century, economic considerations forced libraries to abandon the cataloging of articles in their journals, and they began to rely exclusively on subject bibliographies of the journal literature to provide patrons with access to journal articles. Thus, the print analogs of abstracting and indexing databases are nothing new. However, these printed tools were

selection of journals to index and in their chronological span. For all intents and purposes, if material in a given journal (or even a given issue of a given journal) isn't covered in the abstracting and indexing database, it might as well not exist from the patron's perspective.

Thus, the processes through which the compilers of these A&1 databases select which journals to index, and which articles within these journals should be indexed, are effectively defining the literature in various disciplines. Most library users are unaware of the precise chronological or literature coverage of these databases, or the differences from one database to another (and note that a library generally selects only one database in a given discipline, typically based on a mixture of quality and cost considerations, due to the very high cost of licensing and mounting such a database); indeed most database providers are very vague about even stating their coverage and selection policies, which can be substantially complex. This confusion is compounded by the fact that these A&1 databases evolve over time, and revisit their selection of journals to index, and the indexing policies (i.e. cover to cover indexing, which creates a record for every item that appears in the journal, or selective indexing, which only

---

generally hard to use and were seldom consulted except by librarians and by scholars familiar with their organization.

In the 1960s the organizations that prepared these bibliographies of the journal literature began to employ computers to manage citation databases than were then formatted for print; as the cost of computers began to drop, they made the databases available for online access, either directly or through service bureaus like Dialog or BRS. The first such databases supported relatively well-funded disciplines like the biomedical and health sciences (for example, the MEDLINE database), general science (the Current Contents and Science Citation Index databases), engineering (the INSPEC database), or the business and financial communities (ABI Inform); access to these files was very expensive (sometimes hundreds of dollars per hour) and because of the high costs use of these databases was largely limited to researchers in commercial corporations or occasionally academics with grant support. Universities sometimes offered a very limited amount of subsidized searching (for example, a few searches per year for faculty, or a search or two for doctoral candidates working on their dissertations). Also, because the search systems were not only very costly but also very difficult to use, most searching was performed by trained intermediaries (typically librarians with special training). As a consequence, while these databases were important resources for researchers in commercial settings, they had an extremely limited impact within the academic community.

In the 1980s computing costs dropped to the point where universities could begin to license these databases at flat fees and mount them on local computers for unlimited use by their academic user communities, typically using software that was designed to support access by end users rather than by trained search intermediaries. Usage grew by orders of magnitude; for example, at the University of California, popular databases such as MEDLINE now support in excess of 100,000 searches *per* week by the UC academic community, and the availability of such databases began to have a major impact on university-based research and instructional programs.

The other point that should be emphasized is the very powerful impact of computer-based information retrieval tools in academic libraries. The experience with online **catalogs was that most users of the library found these automated information systems so much more convenient than the card catalogs they replaced that they would typically use the online catalog even if its coverage was less complete than the older card catalog because some material in the card catalog did not yet have machine-readable records that allowed this material to be represented in the online catalog. Similarly, while the printed abstracting and indexing tools were very difficult to use, the online versions of these tools (at least in conjunction with end-user oriented retrieval software typically** used when they are mounted at university libraries) make *the* electronic databases very easy to use, and these databases consequently gain very high user acceptance and quickly begin to serve as the primary-indeed often nearly the sole-means of access to the journal literature.

creates records for certain material in the journal, based on type of material or article content) for selected journals from year to year; just because an A&I database currently covers a given journal at a given level of detail does not mean that it provides historical coverage of that journal, or that it has always covered the journal at the same level of detail.[27] Yet users—at least those in disciplines which still take the published literature seriously, as opposed to disciplines that view the key literature as preprints, technical reports and other electronic publications—tend to regard the coverage of the A&I databases available to them as effectively defining relevant literature in a discipline.

In a very real sense, the challenge facing an author of a scholarly article under the "publish or perish" regime still commonplace in academia for print publication is to get *published;* whether anyone reads the publication is a secondary issue. In the evolving networked information environment, all evidence suggests that it is all too easy for anyone to share their thoughts with the networked community through self-publication. The challenge in the networked environment will not be to make one's writings available, but rather to get people to read them. This will assign an ever greater emphasis on the selection and coverage choices made by abstracting and indexing services, particularly those that are explicitly recognized by scholarly communities because (for example) they are provided by various scholarly societies.

On one hand it seems that this trend is encouraging. Greater importance will be assigned to reviewers and bibliographers of all types. A researcher in a given area may well be willing to pay for the bibliographies of important recent articles provided by major figures in his or her field. Reviewers for journals-currently normally largely

---

[27] Close examination of editorial policies for abstracting and indexing databases indicate that they are very complex and have considerable impact on what information the user locates and how they can locate it. Consider, as one example, a popular database that offers coverage of the parts of the computing literature. Basic records in this database include the author, title, date of publication, subject headings describing the contents of an article and related material. Some, but not all, database records also include abstracts. Some of the journals in the database are indexed "cover to cover", which means that descriptive records for all material *of certain types* appearing in the journals are included in the database—but this may only include news announcements and articles, and not letters to the editor, errata announcements for articles in previous issues, conference announcements and calls for papers, or other materials. Advertising is almost always omitted, even lengthy special advertising sections and the sort of quasi-editorial material like new product announcements that are often found in trade journals. For other types of journals only articles related to computing are included in the database; thus a paper in a journal like *Scientific American* would be included only if it dealt with computing. Since the database vendor incurs a significant additional cost for each abstract that is included in the database, abstracts are only prepared for some of the material, most commonly longer articles. The vendor also offers a supplementary extra cost product that provides full text for some of the material in some of the journals that are covered by database; journals are included primarily based on the ability of the database provider to negotiate an acceptable agreement with the journal publisher for the remarketing of the text of the material in electronic form. Within the journals that are supplied in full text form, the database provider again employs editorial policies to select only specific types of material for inclusion as fulltext, since for most journals the database provider must pay for scanning or rekey boarding of the material and thus again incurs substantial costs for material included. For some types of material there may be fulltext but no abstract. Now, add to this rather complex set of criteria for what is placed in the database the additional complexity that all of the editorial policies just described are subject to continual revision and fine-tuning.

The user of such a database is typically unaware of all of these subtleties. However, searching by subject terms will actually search a different, larger set of articles than those accessible when searching by full text or keywords in the abstracts, and the table of contents of a given journal issue as derived from this database are likely to be somewhat different than the contents of the printed journal.

unrecognized and uncompensated for their labors—may find their evaluations recorded in databases and assigned great importance. Those who edit, filter, and select may play a much more important role in the networked information world. But, at the same time, established arbiters of taste within a given discipline, such as the compilers of abstracting and indexing databases, may have a much greater role in describing the relevant literature of a discipline.[28]

A key question here will be the amount of diversity available. One perspective on the matter extrapolates from the existing compilers of abstracting and indexing databases: these are organizations that attempt to provide systematic and comprehensive coverage of the literature in a discipline. Developing these databases is a costly proposition; the creation of such a database is a major investment by a corporation or other institution. The other perspective uses the network to expand the reach of what has traditionally been interpersonal communication—someone passes an interesting article to a colleague. The individual-based filtering and selection services serve different purposes and in some ways are more valuable to information seekers increasingly pressed for time as they help such information seekers to locate key publications quickly. Here the model is more one of bibliographies and reader's guides, which can be produced for limited areas by a single specialist or a small cadre of experts with a fairly limited investment Of course, individual-based services are more subjective. One of the most attractive points about individually produced bibliographies and reader's guides is that it gives wider voice to major thinkers in a given scholarly discipline-the "geniuses", to use one reviewer's term, can reach beyond their immediate circle of students and colleagues to highlight what they believe to be particularly important works for the broader scholarly community. Both approaches will have their roles.

The entire issue of evaluation of literatures is controversial [White, 1989]. Some librarians and researchers (such as F. W. Lancaster) argue that this is one of the key contributions of librarians and of various reviewing services. Certainly, every library makes evaluations daily as part of its acquisitions decisions, but the often it avoids suggesting that one item it its collection is "better" than another once the evaluation decision leading to acquisition has been made. The argument has also been made that the standard review sources in many disciplines are at best very conservative: they only tend to cover material from certain mainstream publishers (and, indeed, in some cases they are owned by one of the major publishers in the field) and as such tend to reduce diversity and the introduction of innovative new material, in part because librarians at

28 Occasionally, one reads visions of future electronic libraries that include a very intensive reader commentary component. The idea is that readers will attach their reactions and comments to material placed in an electronic library by the primary authors. Effective realizations of such a framework have proved elusive in practice. There are too many readers, with greatly varying levels of expertise and objectivity. While broad-based reader commentary may be a useful thing to incorporate in future electronic libraries, I do not believe that it will replace the role of expert selectors and commentators. It is also worth noting that there are subtle intellectual property problems here. Will the general public be willing to contribute their comments on material for public access? Certainly, some experts will try to make income by providing such commentary; if the public at large emulates this, one has an administrative, legal and accounting nightmare. If the public does not, then one must ask why certain commentators are willing to share their thoughts on a work freely while other commentators are not. Some projects, such as Ted Nelson's XANADU {Nelson, 1988], have attempted to explore the compensation and intellectual property issues implied by a move from published works to a rich web of commentary that surrounds these works.

many institutions, overworked and/or lacking the necessary expertise to make an independent evaluation, will simply use the review sources as purchasing guides. It seems to be that the networked environment will increase diversity in reviewing sources, thought it is not clear to me that many librarians (as opposed to subject matter experts) will step up to the challenge of providing these new bibliographies, abstracting and indexing tools, and reader's guides.

The trend towards having large, costly abstracting and indexing databases define the "core" of a disciplinary literature is of particular concern in conjunction with visions of the future which place professional societies in charge of the canonical literature in a given discipline (see, for example American Physical Society document on the development of a future international physics electronic library [Loken, 1990]); the problem here is that while a given researcher who is out of step with the conventional wisdom in a given field may be able to make his or her thoughts available on the network, it is unlikely that anyone will find them. One can all too easily envision the "establishment" in a given discipline taking control of the definition of the literature in that discipline through the compilation of the de facto standard abstracting and indexing databases in that discipline. To a certain extent, the easy self-publishing that is possible in the networked information environment addresses these concerns, but as indicated earlier the challenge is not to be published but to be read, In cases when tenure and promotion are at issue, there is likely to be no near-term substitute for publication on a prestigious journal; but, when the objective is more communication with one's peers, the question is whether the developing tools for identification and discovery of networked information resources will provide an adequate "safety net" to allow self-published materials to be located and read by those peers.

Another aspect of the role of secondary information services is their role in author evaluation decisions—for example, tenure and promotion decisions for academic authors. Some of the more sophisticated universities are recognizing the potential for subjective bias that may be present in the traditional abstracting and indexing services, and prefer what are allegedly more quantitatively objective secondary services such as the various citation indexing offerings from the Institute for Scientific Information (ISI) such as Science Citation Index [Garfield, 1979]. Citation indices count the number of times that a given publication is cited in the published literature; it is only a short step from these to even more "objective" measures of quality based on the number of times that a given article is accessed (in electronic form, where this number of accesses can easily be computed); this raises fundamental questions about the privacy of searches and the uses to which searches can be put that are discussed in a later section.[29]

The growing power of abstracting and indexing services raises many questions that need to be explored, and places at least a moral responsibility on the abstracting and indexing services to exercise a very high degree of quality control (though the legal liability of such services, as far as I know, has yet to be defined; the general issue of legal liability of information providers is discussed in a later section of this paper).

[29] It should be noted here that citation rates are a somewhat controversial measure of the impact of publications. They are subject to "gaming" in various forms: repeated and extensive self-citation, or the development of tight circles of authors who continually cite each other's works [Pertiz, 1992]. Similar questions will undoubtedly apply to the use of measures based on the number of accesses to articles in the networked environment.

Consider the possible impact of an service that abstracts and indexes only selectively: for all practical purposes, by not including a given article, the service excludes that article from the literature of a discipline and makes it unlikely that researchers in that discipline will subsequently find the article in question. This at least is an editorial judgment, [30] consider the case where by some error an abstracting and indexing service "misses" an issue of a journal and all its contents (perhaps the issue was lost in shipment to the service, or lost by the service during its processing stream), or makes an indexing error which causes a publication to become unretrievable. Such omissions evidently do occur today in some of the major services that are used in contexts such as tenure and promotion decisions.[31] The entire issue of the quality of abstracting and indexing databases is quite complex and subtle; the interested reader might wish to examine the recent series of articles by Peter Jasco on aspects of this topic [Jacso, 1992a; Jacso, 1992b; Jacso, 1993a; Jacso, 1993b; Jacso, 1993c].

As we begin to transition from printed materials to electronic materials we tend to think of abstracting and indexing services (first in print formats, and now as electronic databases) as perhaps the primary means of identifying source materials. In fact, as more and more primary (e.g. full text, or source) material is available in electronic form, new methods of identifying relevant material will come into play based on various forms of automated indexing and full text searching,[32] [Salton, 1988]. This is inevitable for three reasons.

First, the human intellectual effort for abstracting and indexing is costly and the user community cannot afford or is unwilling to pay for people (particularly expensive people with subject expertise) to index everything, particularly in great depth. Even if an abstracting and indexing database is available which covers a given set of material, a library might offer that source material in electronic format but may have chosen not to license the abstracting and indexing database for any number of reasons (for example, because the library only holds a very small proportion of the material that is covered in

30 One of the reviewers of the initial draft of this paper raised a very interesting issues about editorial selectivity: based on the Feist decision, one might argue that a comprehensive, cover-to-cover indexing and abstracting service would have very limited protection under copyright, while a service that was more selective would find that their selectivity would justify stronger copyright protection. Copyright protections may well encourage greater selectivity.

31 In many disciplines there are multiple competing abstracting and indexing services. Publishers have less to fear than individual institutions from errors that are made by a single service; over the broad subscriber base, the multiplicity of services will make it probable that at least one service provides proper access to the publisher's materials. However, given the very high cost of acquiring an abstracting and indexing database, a given library or university will probably select a single supplier from the various alternatives available on the market; thus, for a given university community (within which, for example, a tenure decision is made) a single abstracting and indexing database will dominate.

32 Note that all of the issues raised already about the power of editorial decision making In the compilation of abstracting and indexing databases also apply to fulltext databases, whether created independently of A&I databases or constructed as extensions of these databases. Choices about what to include in fulltext will have to be made; some database producers may not choose to include full text of ail articles that they abstract and index, or more generally of all articles that appear in a given issue of a journal. And new opportunities for editorial bias appear: for example, a given service might exclude full text of articles that are critical of that service's performance or practices. Given that many users will be satisfied with the part of the published literature that is immediately available in full text electronic format, such editorial decisions can have a powerful impact.

the A&I database); in these cases there is no choice but to use information derived from the source material to provide access to it.

Second, while mechanisms based on full text may or may not offer "better" access to material [Blair & Maron, 1985; Tenopir & Ro, 1990], they certainly offer different access which is at least a useful complement to human intellectual indexing,[33] Access based on full text can be an excellent supplement to shallow abstracting and indexing databases (for example, those that provide little or no subject access). Full text access can help to identify documents that mention people, places or things that may not have been sufficiently central to the theme of the work to be recognized by an indexer or abstracter; in this sense they provide much greater depth of access. Further, there is a sense that full text access is less "biased" than abstracting and indexing services in the sense that human judgment does not come into play. Full text access can help users who are having difficulty with specialized controlled vocabularies typically used in subject classification in A&1 databases.[34] In situations where large textual documents are available online, the two techniques may be used together: first, search an abstracting and indexing database to identify relevant documents, then use full text based access techniques to identify relevant parts within these documents.

A third reason why full text based access is coming into wide use is because of the delay inherent in human intellectual indexing. Today, major (and expensive) abstracting and indexing services often run as much as four to six months behind the appearance of source material in print (and recall that the print material itself may be months or even years behind the distribution of preprints or manuscript versions of material with the "invisible college" community). As electronic dissemination of information increases the speed with which material is made accessible, these lags in abstracting and indexing will become increasingly unacceptable to some users of the material-particularly those interested in the most recent material rather than those performing retrospective literature searches. Full text indexing allows access through

---

33 The definition of the quality of a method of providing access to documents is a very complex and somewhat subjective area. In the information retrieval research community measures such as precision, relevance and recall are used+ssentially measuring how many of the relevant documents are retrieved by the access method, how many irrelevant documents are returned along with the relevant ones, and how many relevant documents are missed by the access method. Clearly, performing large scale comparative tests between different methods given these definitions is extremely difficult, since it requires that someone go through the entire database in order to determine the "correct" answer to the queries in order to evaluate the performance of the access methods being tested, because of the great variation in the kinds of queries issued by users (and the great variation in the performance of many access methods from one query to another), and because of the very subjective nature of relevant documents (since even experts do not always agree on whether a given document is relevant to a given query, and the judgment of the experts may still not agree with the judgment of a typical user who is not a subject expert). At the same time, we should recognize that while this is a hard problem, experimental results for various retrieval approaches on a wide range of large databases would be of enormous interest and value.

34 Full text access is also helpful in dealing with the fact that controlled vocabularies grow and change over time as new areas of interest emerge within a discipline and new discoveries and developments occur, but these changes tend lag substantially behind the events that cause them; this is a well-recognized problem with the Library of Congress Subject Heading controlled vocabulary list, for example. In most cases, the cost of updating subject terminology used in existing database records to reflect changes in the terminology is prohibitive; only a few very high quality databases such as the National Library of Medicine's MEDLINE do this. Often, one can find terms used in the abstract or full text of an article long before they become established in the indexing vocabulary of a discipline.

the apparatus of bibliographic organization to occur simultaneously with the act of (electronic) publication. It is also interesting to note in this connection that part of the problem is the size of the literature base that most comprehensive abstracting and indexing services attempt to cover (which goes hand in hand with their lack of evaluative information-they will help you find all the documents on a given subject, but not the three best surveys). If we see the development of large numbers of limited scope, highly selective and highly evaluative citation lists/bibliographies offered by subject experts as proposed elsewhere in this paper, we may find that these specialized lists are also much more timely than the traditional abstracting and indexing services.

Clearly some types of electronic material, such as newsfeeds, will require automated indexing; human indexing will introduce so much delay that much of the time value of the material would be lost. Multimedia information—images, video feeds and the like-present an additional set of issues. Today, we have very limited capabilities to perform useful computer based indexing of multimedia; general image classification is beyond current technical capabilities,[35] though automated transcription of speech (audio, or the audio track of a video segment) may become a production technology within the current decade, and this soundtrack could provide a very valuable access point for video information. Already, today, closed-captioning tracks in video material are being indexed and used to provide access points to broadcast information. And there is technology in experimental use that separates pictures from text in bitmapped images of printed pages [Lesk, 1991], or that attempts to detect scene transitions in video clips .

There are many different full text based retrieval methods [Tenopir & Ro, 1990]. The simplest provide searching for exact words that appear in the text, often with the option of including truncation (match only the beginning of a word), Boolean operators (i.e. AND and OR), and proximity operators (to require that two words appear close to each other) in queries. These full text access methods are easy to understand and predictable, although they often lead to rather poor retrieval results. Much more sophisticated methods have been developed in the information retrieval research community and are now starting to appear in large scale production systems. These range from statistically based methods pioneered by Gerald Salton and his colleagues over the last three decades[36] [Salton, 1988], which are based on frequency of word occurrence along with some very superficial language processing (word stemming) through much more complex techniques that combine statistical analysis with various syntactic and semantic analysis techniques from natural language processing (for example, analysis of parts of speech, identification of proper nouns or noun phrases, or

[35] Certainly there have been great advances in image recognition in very specific problem domains ranging from quality control in manufacturing processes through target identification for smart weapons systems, but more general problems, such as identifying the objects in a picture, remain intractable to the best of my knowledge. Further, really useful classification of images for general purpose retrieval involves a great deal of cultural knowledge as well as simply the ability to identify things: identifying a photograph of the President of the United States shaking hands with the Mayor of New York is far more useful than simply recognizing that a photo depicts two men shaking hands.

[36] In the past few years there have been a number of proposals for more sophisticated and computationally intensive statistically based indexing algorithms, such as the Latent Semantic Indexing techniques developed by Bell Labs [Deerwester, Dumais, Furnas, & Landauer, 1990; Foltz, 1990].

even attempts at actual language understanding). Recently, a major focus on the application of these sophisticated hybrid methods in large production textual databases by the DARPA TIPSTER [Harman, 1992] and TREC [Harman, 1993a; Harman, 1993b] projects has produced some impressive successes and may encourage their transition from research efforts to more broadly deployed systems. The difficulty with all of these sophisticated methods, however, is that their operation is incomprehensible to almost all users. It is very difficult to predict what they will retrieve and what they will ignore. Some critics of these approaches have termed them "information retrieval as magic". These technologies raise very real integrity and access issues in that they work reasonably well often enough to be useful but seldom work perfectly; worse, they fail drastically in a reasonable number of cases. And information seekers not only have no idea what these retrieval systems are doing, but very little sense of when they are or are not working right; and, as they move from one system to another (as will be increasingly common in a networked information environment) they also have no sense of the specific features and idiosyncrasies of a given retrieval system. And, unfortunately, little effort seems to have been invested in researching effective means for these systems to explain and document their processes to their users; such features would help a great deal.

To some extent, these sophisticated "voodoo" retrieval systems have been kept from the general public by groups like librarians who are sufficiently infomation-retrieval literate to recognize the problems and be alarmed by them. The general public won't care; as soon as these developing technologies become effective enough to provide a useful answer most of the time, the public will accept them (and swear at the "stupid computers" in cases where they don't work), unless we see an unprecedented rise in public literacy about information and information retrieval techniques. The unreliability of probabilistic and statistically based retrieval algorithms is today not a problem that the public understands; without such understanding they may well become victim to their limitations simply because they are easier to use than more traditional, deterministic approaches.

6. Access to and Integrity of the Historical and Scholarly Record

One can consider a printed work as knowledge bound at a given time. For example, an encyclopedia published on a certain date represents the common wisdom of society about a number of topics as of some point in time. Indeed, old encyclopedias, obsolete textbooks, out of date subject heading classification guides and other literature represent primary databases for cultural research[37] and for understanding our culture's view of the world at a given time. The scholarly record in any given area, viewed as a series of frozen artifacts narrowly spaced in time can be viewed as such a historical record.

The same issue applies to mass media. The daily, weekly and monthly publications of popular journals provide a nearly continuous chronology of the shifting perceptions of any number of cultural issues. The selection criteria for what is published are themselves a very important part of the cultural record, and represent very definite

371 am indebted to Professor Michael Buckland for illuminating this point.

biases (in some cases, one selects information sources precisely for the benefit of those editorial biases). Further, as information technology has made publishers more **agile,** as we have moved more to broadcast media (where only the present exists, in a real sense, and it is very difficult to go back and look at the media's content at earlier points in time) and as means of monitoring audience response have become more precise and more timely, content can be changed almost continuously in response to audience interests and preferences rather than reflecting a consistent editorial position Indeed, this content shift may take place hour by hour in the popular media: one can envision services such as the Cable News Network (CNN) shifting perspective from one broadcast of the news to the next (every half hour) based on viewer feedback and sensitivities .[38]

In a real sense, electronic information resources invite an Orwellian, a historical view of the world. Consider an electronic encyclopedia that is updated weekly or monthly; entries for countries and political movements are freely replaced. Rather than a series of views of events fixed at specific times, the entire view of the world is now subject to revision every week or two. There is no a priori reason why the implementation of such an electronic encyclopedia must ignore the past, but this is the simplest implementation; overlay the obsolete with the present.

Within the database management system community a concept sometimes termed "time-travel databases" has been developing; these are databases that can be viewed based on their contents as of a given moment [Stonebraker & Kemmitz, 1991]. As the database is updated, older versions of database records are retained, along with information as to when updates were applied and when information is replaced. Records are not actually ever deleted in such databases; rather, an indication is stored that notes that a given record has become invalid as of a given point in time. Such versioning or time travel databases are still at the research stage, however, and most commercial DBMS software does not support the necessary range of functions to allow production implementations of databases that incorporate a historical record of database evolution. Further, even if software becomes available, there are substantial costs in disk storage and retrieval efficiency that must be paid in order to provide historical views of database content. Libraries, facing continued financial pressures, will be hard put to justify investment in these technologies. Yet the ability to retrieve the state of knowledge or belief about a topic at a given point of time is an essential element of the historical scholarly record, and indeed a critical part of the data needed for a wide range of research endeavors.

The shift from sale and copyright law to contract law also raises issues where integrity and access combine in complex ways. Once a library purchased or otherwise obtained a physical artifact (for example, through a donation) that it made part of its collection, this artifact became part of the library's permanent collection. With the replacement of the transfer of artifacts by licensing of electronic information, it becomes much more difficult for a library to maintain early editions, erroneous distributions and other

[38] Continuous news broadcast services such as CNN currently modify about 6-8 minutes of their coverage from one cycle of the news to the next, dropping stories, adding stories, or making editing changes to stories that are repeated from one hour to the next (with these editing changes not necessarily being necessitated by new news developments).

materials that may be a part of the historical record which the publisher of a given information resource is not necessarily eager to have generally available. One need not assign any malice to the publisher wishing to withdraw out-of-date versions of their publications from circulation; the publisher may be doing this with the best motives, such as ensuring quality control. A pharmaceutical database publisher may want to ensure that incorrect or obsolete information (which may, indeed, be dangerous—for example inaccurate dosage data) is corrected promptly and comprehensively. More generally, a the publisher of an electronic newsletter may simply want to ensure that the corpus of published material is as accurate as possible; there is an inherent conflict between quality of a published corpus and the accuracy of the electronic publication as a historical record. The integrity of the historical record becomes far more subject to the desires of the publisher.

Earlier in this paper, copyright was identified as a potential barrier to access in the electronic environment. In the context of integrity, however, it can serve a very valuable purpose for authors by providing a basis for the author to ensure the integrity of his or her words over time, and preventing later "amendments" or "corrections" to published works. The right to make changes, like other rights (such as republication or translation rights) is subject to negotiation between author and publisher.

Of course, by the same token, one can imagine situations where a publisher (for example, a government or some other entity) uses its license control over material to effectively rewrite the historical record; certain material is simply declared "inoperative" and removed from circulation.[39] This is another illustration of the extraordinarily strong position of publishers, authors and other rights holders in the electronic information environment and the loss of public policy control of the balance between the rights of creators (or rights holders) and the public.[40] Ultimately, there may well have to be a rethinking of the definitions and meaning of publication; in the print world there is a strong sense that once something is published, some copies are distributed and available to the public permanently. Even in cases where a lawsuit is successfully brought against a publisher for one reason or another, while a result of the judgment may be that the publisher ceases to sell the work and destroys existing stock, there is really no practical way to recall copies already sold. Publication in the print world is generally viewed as an irreversible act;[41] at least under some definitions of

39 While slightly outside of the main focus of this paper, one area that I find particularly interesting and troublesome in this context is control of the news. The primary record of historical events is copyrighted material owned by newspapers and the broadcast media. As the various trends discussed in this paper lead to a situation where less and less of this material is held by libraries, it raises the specter of situations where for whatever reasons the primary historical record of events in a given area might well become inaccessible to researchers.

40 Copyright is not the only issue here, however. For example, an executive order was signed during the Reagan administration that permitted the government to reclassify previously declassified material in cases where they were able to regain control of all copies of the declassified document. In a print environment this is quite difficult; in an electronic environment it would be much easier.

41 Because of this irreversible nature of the act of publication, the scientific community has had to develop the practice of withdrawing a previously published paper that is later been found to be erroneous, for example; this is accomplished by printing a notice in a subsequent issue of the journal. This is not necessarily very effective, since a reader of the withdrawn paper may be unaware of its status. In an electronic environment, it is interesting to speculate how a withdrawn paper would be handled. Would it continue to be distributed, but bearing a prominent notice that the author has subsequently withdrawn it, or

"publication" this is not the case in the electronic environment. The new electronic environment is likely to create a great demand for what are perceived to be neutral parties to maintain the historical and cultural record (and I believe that most people view libraries as falling into this category), as well as various forms of audit trails so that revisions of the "record" can be tracked and evaluated.

The ability of a library to acquire access to data rather than copies of information resources also threatens the historical record of scholarship and culture. One vision of electronic information resources calls for publishers to mount databases of journal articles on the network, rather than supplying copies of the information to libraries that subscribe the these journals. In such a world traditional journal subscriptions are replaced by contractual arrangements that allow libraries to retrieve articles from the publisher provided servers under various terms (pay per view, or unlimited access to articles from the journal through a "subscription" price, for example). This may represent an economy for libraries, in that they do not have to receive journal issues, and they do not have to pay for local storage space to house these journals. But what happens when a publisher decides that a given journal (or specific back issues of that journal, or even specific journal articles that appeared in the journal) are no longer being used enough to make it profitable to provide access to the journal on the network? Or, perhaps, the publisher goes bankrupt, or is acquired by another publisher, or becomes entangled in litigation? In such cases it is quite possible that the publisher will cease to provide access to some or all of the contents of journals without notice, and without any recourse by the library community; the back issues simply become unavailable, for example. In the old print world, if a library somewhere held these back issues, they would still be available to patrons through interlibrary loan, but in the new electronic environment, unless some library had made local copies of the material (thus losing out on the economies that make electronic distribution of the matetial attractive) this material could be lost to the user community for all time.

Copyright law may again provide a useful tool for ensuring preservation of the scholarly record. Historically, deposit of a copy of a work with the Library of Congress has been a requirement for establishing copyright protection; while current copyright law has removed this requirement, to a great extent, a return to such deposit requirements could help to ensure the long-term accessibility of electronic material.

In the print world archival access to material was the responsibility of the library and archival communities. In a world of licensed access to electronic information, libraries cannot unilaterally continue to accept and discharge this responsibility. It may well be that the networked information environment will call for a new compact of responsible behavior between publishers and the library community, in which publishers make a

---

would distribution cease? In the OCLC/AAAS ***Current* Clinical Trials** electronic journal, the reader of any article that has had subsequent corrections or comments receives a very prominent warning that such supplementary information exists; however, it is in some sense **easy** for ***Current Clinical Trials*** to make such linkages visible to the user, since the journal is not simply distributed as content but includes a integral OCLC-supplied viewing **interface. Unless the historical record is actually altered in an electronic journal (at least to the extent of indicating as a note in an article that a correction or withdrawal notice was later issued, and when) electronic journals distributed as pure content (without a user interface to make such** links) **are** likely to offer only the same weak ability to notify users of subsequent corrections that characterize the print publishing world.

copy of their material available to some organization serving (and governed by) the library community so that the library community can assure itself of continued availability of material. Or a publisher might agree that if it removes material from availability on the network, it will offer this material to some access provider of last resort that is financed and governed by the library community (perhaps a network analog of the Center for Research Libraries for example). But the problem here is that while it is reasonably straightfoward to find solutions in an environment of cooperation between libraries and publishers in which all parties behave responsibly, there is the constant threat of irresponsible behavior on the part of publishers, or of external, uncontrollable events giving rise to the loss of key parts of the scholarly record.[42] National attention to the role of national libraries or other organizations in ensuring the preservation of, and access to, the scholarly record, is of vital importance in gaining the confidence of the user community in abandoning printed formats for electronic ones.

It should also be noted that there is another, more crass, issue that is raised by the transition to a networked information environment in which publishers are the primary providers of their inventory. Print is an inherently distributed medium, whereas in the electronic environment a technically inept publisher might stand to lose their intellectual property holdings through various types of catastrophe like fire, earthquake, or corruption of a network server by computer hackers. While the user community would not loose the rights to their material, practical access to this material might well become permanently lost.[43] From a business point of view it might mean that the publisher went bankrupt, but from the broader perspective of the scholarly community, it means that the material is lost and is no longer a part of the scholarly record. Given the numerous relatively small publishers, such as professional societies that issue one or two journals, loss of information due to failures of the publisher to adequately back up or protect their material should be viewed as a very real issue.

Natural disasters and business failures are not the only issues. As libraries move from providing access to their own local physical collections to a set of networked resources international issues must also be considered, for example. One can readily imagine situations where a national library in a foreign country provided access to the majority of the literature related to that nation, until suddenly some international political problem (an obscenity dispute, a war, a change of government, or whatever) caused that national library to cease providing the information in question. Even if there was no central point of control such as a national library access to information provided in one nation could be cut off for other nations by government action. Access may not just be interrupted; more subtle changes are possible. Imagine a fundamentalist religious government taking power in some nation; they might order the destruction of some

---

42 The issue of the scholarly record in electronic form should not be viewed in an entirely negative light. Today, in print, retractions and corrections probably rarely reach those who read the original article. In an electronic environment where we can track who has read (or downloaded) a given paper, the possibilities for disseminating retractions or corrections to the readers who most need to be aware of them is greatly improved.

43 Some publishers have argued that downloading in the Internet environment implies that there is always likely to be some copy of a publication stored on some individual's workstation, and that in this sense electronic publication on the Internet is also an irrevocable act. But, I would suggest that there is a great difference between continued access to material by some random member of the scholarly community and continued access by an institutional agency (such as a library).

materials (or at least the removal of these materials from an electronic information archive) and in other cases change indexing terms in such a way as to distort the functioning of traditional bibliographic access apparatus.

Access to the historical record is not merely an issue of ensuring access to the primary material. In a very real sense, as already discussed, the coverage of the abstracting and indexing services defines the literature of a discipline for many information seekers. But, in almost all cases, abstracting and indexing services began creating computer-processable records in the late 1960s or 1970s. Except for monographs covered in library online catalogs the literature prior to those dates is inaccessible through computer-based retrieval tools, and, for all intents and purposes, might not exist in the mind of many library users. In some fields, particularly the humanities and some social sciences, programs will have to be established to make the remainder of the scholarly record accessible in electronic form and thus place it on an equal basis with the recent publications that are abstracted and indexed by electronic databases. Considerable work is needed in establishing priorities for investment in both the abstracting and indexing of older print material and the conversion of the source material itself to electronic form; these priorities must consider both the mandates of preservation (creating electronic versions of material that is currently deteriorating because it was printed on acid paper, for example) and the programmatic demands of the scholarly communities that will use the materials.

7. The Impact of Micropublishing, Narrowcasting, and Information Feeds

Publishing is becoming increasingly fragmented under the dual pressures of advertisers and readers. For advertisers, a given publication venue becomes increasingly attractive to the extent that it can offer the advertiser a very specifically focused, homogeneous readership-families in the San Francisco bay area that make over $50,000 per year, that are employed by service industries, that collect stamps, and that are shopping for their first house. From the perspective of the reader, overwhelmed by the ever increasing flood of published information, publications that are highly specific to the subscribers interests are of much greater value than more general periodicals. Improvements in the technologies of composition and publication have facilitated this fragmentation to attract readers and advertisers. It is quite common today to see mass market periodicals published in a large number of regional and industry-specific editions; newspapers now offer a wide range of regional editions. Indeed, today's regional editions are composed out of common article databases, but the interest templates are established by region, industry or other criteria rather than individually. The inexorable march of technology seems to point towards ever greater specialization to the audience, ultimately all the way down to the individual reader obtaining a custom product; this trend is manifested in developments that range from the experiments conducted at the MIT Media Lab in the composition of "personal newspapers" based on filters to newsfeeds [Brand, 1987]. Apple's experimental prototype Rosebud system[44] which allows users to define and operate software agents under the metaphor of "reporters" that cover specific information sources [Kahle, Morris, Goldman, Erickson, &

[44] The information access model pioneered by the Rosebud project has recently resurfaced in the commercial AppleSearch product.

Curran, 1992b] or cable television narrowcasts (for example, the slotting of five minutes of local news into a CNN broadcast every hour). Other variations on this model are already coming into large scale use within corporations; for example, in financial and securities markets that are highly sensitive to news, real time news feeds are being licensed for distribution over corporate local area networks, and these news feeds are filtered to allow near-instantaneous delivery of relevant stories to the appropriate individuals very rapidly [Belkin & Croft, 1992; Goldberg, Nichols, Oki, & Terry, 1992; Marshak, 1990].

These trends create enormous problems for libraries, which will only become worse as more information is cast into electronic forms suitable for personalized retrieval. In the case of print, it is becoming very difficult to track what has appeared in the published literature, and who might have been aware of its appearance. Different libraries hold different editions of newspapers and magazines. Abstracting and indexing services do not typically cover all of the various regional editions of a publication; rather, they select one (sometimes without much consideration, and without being very clear about which edition they have selected) for indexing. As we move beyond print editions to the databases from which the specialized print editions (and future personalized extractions) are generated, the situation changes again. It now becomes possible for a library to obtain or purchase access to what is in essence the composite intellectual contents of all of the various editions, but the researcher coming to these databases years later may well loose any understanding of how items from these databases were selected (or, sometimes more to the point, not selected) and presented to any given audience or what the impact might have been on readers. The selection of available materials into a given edition represent a value judgment about the importance of specific subjects in time and space, and the record of this judgment is of critical importance to researchers. The key questions, as we attempt to mine these databases for research, will be what material a given reader of a given print edition learned from that print edition, and what material would likely have been selected from a given database at a given time by someone filtering the database with a given interest profile. To make matters worse, the filtering typically occurs (at least today) close to the end user rather than the publisher, and there is a great diversity of filtering tools ranging from rather simplistic keyword matching all the way through sophisticated research tools that perform semantic and linguistic analysis to identify material that may be relevant.

There is also a serious problem with the accuracy of citations while we remain in a transitional stage between print and electronic databases. n the fully electronic world, one might simply reference article ID X (added on such and such a date) in the Wall *Street Journal* article database. But today, this would most likely be referenced by its title, and the date and issue number of the issue that the article appeared in, without providing the critical bit of additional information—was it the East Coast, West Coast, European, or Far Eastern edition?

The recombinant nature of materials that are maintained in electronic databases but presented to the reading public through specialized printed vehicles is not limited to newspapers and magazines. McGraw-Hill has embarked on a program called PRIMUS in which they supply databases of articles that can be combined into course readers, with the inclusion of optional commentary or additional material by the compiler of the

course reader. Here again it is becoming increasingly difficult to determine the provenance of material, its timeliness, or the conclusions that the reader might have drawn from it based upon context.

The ready availability of desktop publishing (now being combined with transmission via fax or electronic mail across telecommunications facilities) has led to an incredible proliferation of very narrow audience newsletters; many of these have very high time value, very high subscription costs, and sometimes rather high impact on the communities they serve. Yet to those outside of these select subscriber communities, these publications are almost invisible. They are not available in libraries, except occasionally for specialized corporate libraries (which often will not circulate the material through interlibrary loan), and material in them is not indexed in commercial abstracting and indexing services that researchers would use to try to obtain information about a subject, or to reconstruct events after the fact.

A surprising parallel can be drawn to the infamous "gray literature" of technical reports in some areas of science and engineering; these are poorly indexed, difficult to obtain documents that play a key role in circulating information within an insider community, but are not readily accessible either through libraries or through the bibliographic control and access apparatus of abstracting and indexing databases. Yet, for an active researcher, they are invaluable, and the information propagated through such technical reports can have an extraordinary impact on a scholarly community. The contents of the technical reports will appear in the traditional, "archival" published literature only years later in many cases, and often in an edited (reduced) form due to page limitations in the archival journals. In some disciplines, important work that is described in technical reports does not always make its way to the mainstream literature, with the authors simply viewing it as too much trouble to manage a manuscript through formal submission, peer review and publication when the material is already somewhat stale and dated to the authors. In some academic disciplines, it is probably not too strong a statement to argue that were it not for the need to publish in traditional archival print venues for tenure and promotion purposes, most of the effort would go into producing material that would be part of the gray literature. Increasingly, in part as a recognition of the importance of such material in scholarly communication and the diffusion of research, such reports are being made accessible through the Internet from file transfer (FTP) or GOPHER servers, most often through initiatives at the departmental level. Of course, this leads to an essential literature that is very hard to obtain access to, particularly for those who are not comfortable with the relevant technologies, and the long term accessibility of this literature is questionable, since there is today little institutional commitment to ensure continued availability of it; partially in a response to these trends ARPA is funding a major project to improve access to technical reports in computer science through the network. 45

[45] If anything, the scope of the gray literature is becoming larger and more confusing as we move into an electronic environment when anyone can make material directly available for anonymous FTP from a personal workstation or departmental storage server. Not only **are** members of **the academic community mounting technical reports but in some cases also the text of material that they have published,** often in violation of the transfer of copyright agreements that they have executed with the print publishers of the work as a condition of publication. Thus far, this has occurred only at the level of individual authors and not on an institutional basis and to the best of my knowledge the print publishers have not attempted to enforce their control of the rights with these individuals, perhaps feeling that it is not worth the expense or fearing that such an action would mobilize the community of academic authors to pay more serious

Over the past few years, the research library community has begun to devote resources to improving access to the technical report literature as part of a recognition of its importance. Several major research libraries have developed extensive technical report collections, and Stanford University has developed an extensive bibliographic database of technical reports which is available on the Internet. Yet the newsletter literature remains almost inaccessible to the academic community and the general public.

While somewhat outside of the primary focus of this paper, mention should also be made of real-time information feeds that go beyond human-generated intellectual property such as newswires. Other types of real time information feeds are beginning to appear on the network; these information feeds, just like newswires, can be analyzed by personal workstations to sift out events of interest. If ever knowledge was power, these applications illustrate the axiom; in some cases, the advantage of timely knowledge goes beyond financial or professional gain to matters that are literally life and death. The ownership, access to, and integrity of these resources is of potentially critical importance, and may in future become a central public policy issue. Consider the following examples of information feeds that might pass over the network:

- Newswires. These are clearly intellectual property (representing reporting of news events) yet timely access to news may permit an individual to obtain substantial financial or professional gain. Related to newswire feeds will be a wide range of real-time audio and video telemetry of current events of interest, similar to what is sometimes found on the C-SPAN cable television network or carried on current experimental services such as Internet Talk Radio. The intellectual property rights involved in these audio and video feeds is far less clear; in some cases the events are public (for example, Congressional hearings) and in many cases they are simply captured through fixed video cameras and microphones, thus involving little creative work.

attention to copyright transfer agreement terms. Often, it is extremely hard to tell precisely how the electronic version available from the author is related to the printed publication; in some cases, the electronic version does not even provide a citation to where the work appeared in print. And, of course, the electronic version is often available months or even years prior to the availability of the published printed paper. Another interesting area of expansion of the gray literature is doctoral dissertations. Most universities use University Microfilms Inc. (UMI) as a means of ensuring general availability of theses, and request (sometimes essentially require) doctoral candidates to file their dissertations with UMI as part of the process of filing the dissertation. The UMI agreement gives the author non-exclusive print rights, including the right to publish articles or books based on the thesis, but reserves *exclusive* electronic distribution rights to UMI. Yet theses, or versions of theses that are issued as technical reports by academic departments are starting to appear for public access as well. Again, it is difficult to tell if these are precisely the same as the formal filed theses; the print versions that one obtains from UMI normally include the signed cover sheet and are identical to the versions filed with the university. H should also be noted that while the incidents discussed here of making dissertations (or variants of dissertations) available electronically are isolated individual cases, there are also institutional level discussions going on between a number of universities and UMI under the auspices of the Coalition for Networked Information about how to make dissertations available electronically through the network and what (if any) UMI'S role should be in such an enterprise. These discussions might well result in a situation where UMI serves as an electronic publisher for dissertations on behalf of the university community (in print, UMI is a very cost effective means of providing access to dissertations, with a typical dissertation only costing about $35); alternatively, it might result in universities negotiating for changes to the UMI copyright transfer agreement and mounting dissertations themselves.

- Stock or commodities prices, or other financial data (for example, purchases of rare coins). It is much less clear that this is intellectual property rather than brute facts, but again timely access to and analysis of such information can offer an individual substantial advantage. Currently, one can obtain subscriptions to the stock market trading “ticker” (at considerable expense).

- Time synchronization information. The US Government, as a public service, offers very precise time information suitable for setting clocks in a distributed computing environment so that multiple computers can operate on a common time scale. (These efforts are supported by NIST, based on programs that date back to when NIST was the National Bureau of Standards). A related service is the Global Positioning System (GPS) satellite network, which allows a user to locate the position of a receiver with a tremendous degree of accuracy. GPS was deployed to support military applications, but has extensive civilian uses that include search and rescue, navigation (for planes, boats, and even potentially automobiles or individuals on foot). Currently, such information is publicly available, and is of substantial value; will private commercials services have a role here in future?

- Weather information. Much of this is collected from sensors (obseving stations, satellites, etc.) financed through public funds. Yet subscriptions to weather information are in many cases through private information services (though the University of Michigan, for example, makes information derived from such private services available to the general public across the Internet through the Weather Underground service). In some cases this information is merely interesting (e.g. whether it will rain today); in other cases it is of financial value (implications for commodities prices). But in a few cases—tornado watches, flood warnings, and the like—it is a matter of great importance to recipients, if they can obtain the information in a timely fashion and act upon it.

- Surveillance data. We are awash in digital imagery ranging from data generated from satellites (either government or commercial, such as the French SPOT system) through output of surveillance cameras in our workplaces and homes. Again, awareness of such information can have value ranging from insight into commodities or securities investment through convenience-traffic congestion information, for example—to personal safety.

- Earthquake warnings. Earthquake shock waves propagate from an epicenter rather slowly compared to the near speed-of-light propagation of information across copper or fiber optic trunks. Some states, such as California, have sensor networks deployed which could, at least in theory, propagate information about the occurrence and epicenter location of earthquakes across the Internet in such a way that locations that will be hit by a major earthquake might obtain as much as 60 to 90 seconds advance notice of the event. This is information which, if identified and acted upon in a timely fashion, could save not only a great deal of money (for example, by “safeing” processes ranging from parking heads on disk drives or stopping elevators in a graceful fashion at floors through shutting down industrial operations such as chemical refineries) but could also save lives. It is unclear who would own such information, or could provide it to the network.

Ownership is not the only issue with such information or other sensor feeds; integrity is equally important. It is vital that such information be not only correct but also accurate and authenticatable. The notion of someone "simulating" a major earthquake through the network, for example, is clearly unacceptable. In this connection, it is interesting to note that most of the existing cryptography based authentication technology could be quite problematic in these applications, unless considerable care is applied to address scaling issues; for example, imagine every workstation in Northern California simultaneously trying to obtain the public key of the earthquake information server to validate an earthquake warning, throwing the entire Internet into overload at precisely the time that smooth operation is most needed.[46]

8. Privacy issues in access in the networked information environment

Confidentiality and Anonymity of Access

Within the library community, confidentiality of collection use information is a well established principle. Library practice, as defined by the American Library Association, defines circulation records as highly private, to be revealed only under a court order—if then. Indeed, practice goes further-typically a library will only store the information that a given patron has borrowed a given book during the period while the book is on loan; once returned, only statistical information based on the patron's statistical categories is retained for subsequent management analysis. Most libraries, even under court order, can provide little or no history about the books borrowed by a given patron. Through statistical analysis, they may be able (if their circulation system is well designed and well implemented) to provide lists of the hundred most popular (in the sense of most frequently borrowed) books, or the ten books borrowed most often by high school students in the past year. It is also usually possible to find out how often a given book has circulated, or how many items a given patron has borrowed in the past year. In fact, such information is very important for the tuning of collection development strategies, for deacquistions decisions, and for overall management and budgetary planning.

Similar principles about privacy in the networked environment are far less clear; there is no general consensus about norms of behavior. Most users have a tendency to assume that their privacy is protected-either by legislation or generally accepted practice-to a greater extent than it probably really is, perhaps making this assumption by analogy to library practices and other situations, such as video rental records.

---

46 Another example of the use of the network to provide information to control machines is provided by the "smart power" technologies that are under discussion in projects such as the Blacksburg Electronic Village effort in Virginia [Bull, Hill, Guyre, & Sigmon, 1991]. The basic idea here is that under heavy load the power company must purchase additional power from other power companies on the national electrical grid at very high prices and it is very advantageous to them to be able to reduce loading during those times; additionally, their pricing, particularly to residential customers, does not let them recover these premium costs directly during periods of very heavy load. Instead, residential costs are to some extent averaged over long periods of time in setting rates. The proposal is that consumers would install smart appliances and controls (thermostats, refrigerators, air conditioners, etc.) that would be connected to the network. During periods of heavy power demand, the power companies would broadcast alerts through the network and these devices would reduce power consumption temporarily. Apparently, preliminary studies on the Blacksburg project have suggested that if the power company actually paid for smart thermostats (assuming that the network infrastructure was in place) they would recover their costs within two years.

Service providers, including libraries operating online catalogs and institutions supporting anonymous FTP archives, have little legal or policy guidance and view the situation with a considerable degree of unease.[47]

If one considers libraries, which at least have a historical policy context that might help them to develop policies for privacy in access to new online information services, one finds a variety of practices and motivations behind them. Many online catalogs provide anonymous access simply because it is easier than having to maintain user files for a large, and, in a university, rapidly changing user community, and not because of any policy commitment to the right of anonymous access to the online catalog (as distinct from a possible policy position on the right to privacy of searches; in other words, the library is saying that it will protect privacy, perhaps, but not providing the absolute guarantee of privacy that anonymous access gives to the user). Some institutions controlled access simply as a means of regulating resource utilization; these controls sometimes required users to identify themselves through user IDs and in other cases preserved anonymity by using controls such as originating network address to determine whether a user had access. As online catalogs have expanded to include abstracting and indexing databases and other electronic resources licensed to specific user communities, it has become necessary for many systems to implement some type of user identification mechanism in order to control access to these licensed resources in accordance with the license agreements. A few institutions, such as the University of California, have developed approaches to user identification that provide for patron anonymity, but many have simply gone to a user ID mechanism, often based upon the library card number. To some extent the questions about accommodating anonymous access tie back to the library's overall priorities; in a period of intense pressure on budgets and library resources, many libraries are articulating strategies that place priority on serving their primary clientele (for example, members of a given university community) and provide access to other users on a lower-priority basis.[48] Members of the primary clientele will typically be registered with the library and this registration process provides them with user IDs that can be used for authentication (thus shifting the issue from guaranteed confidentiality through anonymity to policy confidentiality provided by the library). As online catalogs grow into ever richer and more complex mixtures of public and restricted access licensed information resources, it is much simpler to abandon the attempt to provide anonymous access when feasible and move towards a uniform authenticated access model, which is less problematic for the

[47] T. a great extent library patrons are protected more by practice than by law; libraries do not collect information such as what books a patron has borrowed beyond the point that he or she returns them. There may be some statistical information used for collection development that links demographic characteristics of patrons to borrowing records, but the old sheet of paper or card in the back of the book in which the names of those people who have borrowed the book over the years has largely been eliminated by libraries, at least in part in response to growing concerns about patron privacy. In the electronic environment, we may see a clash of cultures; telephone companies, for example, typically gather and retain very detailed records of who each customer has talked to and when; these **are** easily accessible with a court order.

48 This is also being done for networked information resources; for example, some FTP sites limit access during the day by users outside of a specific set of network addresses that are viewed as defining the primary user community, or the limit the amount of anonymous traffic to ensure that resources are available to serve the primary clientele.

primary clientele than to other outside users who wish to make occasional, casual use of the library's information system through the network.

There are other reasons why online catalog designers are moving towards (at least optionally) authenticated access as online catalogs become more sophisticated [Lynch, 1992]. This is needed for current awareness services that electronically mail notification of the arrival of interesting new materials to users, for intelligent interfaces that track a user's history with the system and his or her preferences, or that tailor the dialog to the user's familiarity with the system based on how long and how often he or she uses it. Again, it is certainly possible to support both authenticated and anonymous access modes, and even to permit users to store preference files external to the library information system, importing them when they start an anonymous session and exporting them again at the close of the session, but all of these options add considerable complexity to the system design, the cost of which is certainly subject to question, particularly in the absence of any policy or community consensus that underscores the importance of offering an anonymous access mode.

Matters are complicated by several conflicting sets of demands on service providers. Indeed, this conflict goes beyond operational needs to a basic conflict of values between the library community's tradition of free access to information and the computer community's emphasis on tracking, auditibility and accountability. Computer and network security practices stress the importance of audit trails and other monitoring tools to protect systems from intruders, and system security and integrity are major issues for any service provider on the Internet. In fact, there seems to be consensus among many of the regional network service providers that anonymous access to the Internet is unacceptable (for example, providing access to terminal servers that can TELNET to any Internet host without first identifying the user at the terminal server so that attempts to break into system can be tracked back to an individual at that institution—but note here that there is no requirement that the information be propagated outwards from the source terminal server, only that it be maintained so that by cooperation among organizations a trail can be defined back to an account at the first institution). Certainly, there are many systems that permit anonymous incoming access, but in order to satisfy these restrictions they limit access going back out to the network to specific, limited sets of hosts that have agreed to permit anonymous incoming access. For applications where there is recharge for information access, careful tracking of users is needed to allow discrimination among user groups. This question of anonymous access to the network has sparked bitter arguments between the library community and the networking community, as many libraries view themselves as potential access points to the Internet, and at least some libraries have taken the position that they should not have to require users to identify themselves in order to access resources on the net that are willing to accept these incoming connections. It seems likely that as "public-access" resources on the network multiply, and particularly as federal, state[49] and local government information becomes more

[49] In California, Assembly Bill 1624 is currently under consideration, which, if adopted, would require that various legislative information be made available at little or no cost to the general public through the Internet. One serious proposal by some members of the legislative staff is that the identity of those members of the public requesting this information be tracked for various reasons.

commonplace that the conflict between security and the right to anonymous access will continue to be troublesome.

Many of the information services being offered on the Internet are viewed as somewhat experimental; indeed, we are all still learning how to build user friendly and effective information retrieval and navigation tools, and analysis of user sessions is a key tool in improving the quality of such systems, as well as more routine tuning and capacity planning efforts that are part of the operation of any large scale service. Finally, it is important to recognize that not all information providers on the Internet are institutional; for example, it is quite common to find academic departments, research groups or even individual faculty members setting up anonymous FTP directories to permit people to obtain copies of their papers and research reports. They view this as not much different than responding to postcards asking for offprints or orders for technical reports, and retain a natural curiosity about who is reading their work (which was evident in the days when they responded to requests for printed copies).

Ironically, part of the problem is the development of the distributed computing infrastructure. Ten years ago, when online catalogs were initially being deployed by most libraries, access was primarily from within the library, or perhaps from a few large timesharing hosts on a university campus; if the library was recording searches, it would typically only know that a given search came from terminal 43 within the library or from machine X (which might have 500 registered users). The identity of individual users accessing resources on the network was effectively hidden behind these large multi-user timeshared hosts, and, while a given network resource might require a user to identify him or herself in order to use that resource, the user was aware when such a request was issued by the remote system-one was asked to log in, or provide a password. Very little information about the identity of individuals accessing a remote service could be determined autonomously by the remote service; if the service offered anonymous access (that is, it did not ask for information from the user accessing it) then the user could have a reasonable degree of confidence that access really was anonymous (barring collusion between the user's local host and the remote host; statistical analysis of who was logged onto the user's local timeshared host in comparison to when a remote service was accessed from that timeshared host could, over time, probably allow a sufficiently interested analyst to trace accesses back to individuals, but such activities are rare, and most users view them as too much trouble to represent a serious threat to anonymous access). As we have migrated to a world of personal workstations, the origin address for a search (or a request to fetch a copy of an electronic document) is linked to a specific host address, and increasingly this host, which is now a workstation, is now in the service of a single master. In the new networked environment, the source of a query or a request suddenly provides a great deal of information about the identity of the individual issuing that query or request. This should not be narrowly viewed as a matter of personal privacy; in fact, in the network environment, it is often hard to identify an individual but easy to identify the individual's organizational affiliation by the network number in the incoming Internet address. While people outside of organization X may find it hard to determine that a given address is person Y's workstation, everybody can tell that the access has come from organization X. This may be a matter of competitive intelligence rather than personal privacy.

Certainly there are technological solutions to the problem of one's address revealing one's identity. The simplest is to carry forward the time honored method of mail drops (post office boxes, or the mail forwarding services that various newspapers have long offered in conjunction with personal advertising). Electronic mail based dating services offering such anonymity through the agency of a mutually trusted third party are already operating on the network; a similar service could easily be set up for TELNET. But, as the number of protocols multiply and distributed system architectures become more complex, the development of general purpose anonymity services will become quite problematic. Further, one must wonder whether the vast majority of users will recognize when their use might be appropriate; the example of dating services is a good one since it is simply a recreation of existing practice in the electronic environment in a fairly direct way, and consequently its use in the electronic environment is appealing to the same people who would likely have used it in a non-electronic world. Whether users will recognize the new risks introduced by the development of new electronic information services, or the redesign of old services for the electronic environment remains an open question.

We are only beginning to explore the challenges that distributed computing raises for individual privacy in the context of "anonymous" remote terminal access becoming increasingly easy to trace back to an individual as more users use their own personal workstations rather than large timeshared hosts. At least in the remote terminal emulation environment—be it TELNET or more modern X Window system based applications-the user employs widely available, well documented, industry standard utility software that is written according to publicly available specifications and which can be used with a very wide range of remote services. Often, software to implement protocols like TELNET and the X Window system is available from multiple sources for a given platform (both commercial software suppliers and public domain or "shareware" sources). While there are some true distributed client-server protocols that are well documented national or international standards, such as the Z39.50 information retrieval protocol, and these protocols are implemented in multiple client software applications that can again be used with a wide variety of remote servers, in the developing client-server oriented distributed environment we will see providers of information services implementing custom software clients. These clients will be distributed to users in executable form only; they will employ proprietary protocols, and will be needed to obtain access to specific information servers. In essence, the user of such a service is expected to execute a program of largely unknown function *which typically has full access to the files on his or her personal workstation, given the current state of the* art *in the operating system* software *that runs on most of these workstations,* and which opens and uses a communications channel to a remote service as part of its normal, expected behavior. This is already the case in some commercial services, such as Prodigy [Burke, 1991].

The opportunities for collection of information are endless; for example, such client software might upload a list of what software the user has installed on his or her hard disk,[50] or the list of USENET newsgroups to which the user is subscribed.[51] Unlike

[50] Lists of software installed on machines is useful not only for marketplace research or marketing demographics (for example, to identify people who might be interested in add-on software to an existing product or in competing products) but for other purposes like identifying illegal copies of software: a

general purpose utility software (for example a TELNET-based terminal emulator), the covert information collection activities of specialized client software may be very difficult to identify and monitor,[52] and while very sophisticated users or institutions may be able to address this problem legally with the supplier of the service (and the client software), most users will likely remain unaware that the problem even exists. We may see organizations giving away client software and access to certain remote services through that client software just to be able to get users to run the client and unwittingly export information that the service provider can use directly or resell to others. We may find a direct contradiction between realization of the distributed computing potentials of the Internet and individual user privacy.

There is another interesting relationship among pricing, privacy and the capabilities of systems supported by information providers in the distributed computing environment. Currently, information providers frequently charge based on the amount of information that is exported from their systems; they offer filtering tools of varying degrees of sophistication. On a purely technical basis, some users of some system choose to do fairly unselective extractions from the information providers and then do ranking and filtering on their local machines; this has the effect of preserving some privacy (since the fine details of what the user is interested in are not conveyed to the information provider) but also tends to run up a large bill since the information provider assumes that everything that is exported is of value to the user and will probably actually be examined by the user, rather than filtered by a computer program running on the user's machine. As information provider capabilities improve, the decision as to how much information to give the information provider in order to permit the provider to perform filtering will likely be based in part on how specifically the user is willing to reveal his or her interests to the information provider; privacy (gained by the method of asking vague questions) will have a price. Balancing this, however, we should note that the trends in technology are towards user clients that act as integrators for *multiple* information providers, not just one providers, and such an integration function obviously cannot be pushed outwards to the providers, since no individual provider has the full range of information necessary to do the ranking and filtering of information from multiple sources.

---

company making multiple products could use one to scan for the presence of copies of others, and then check its registered user files.

51 The suggestion that a local client could exploit information about a user's subscriptions to USENET newsgroups is due to Simon Spero, although he proposed it in the context of client software using this as hints in developing a user profile which could be used to help tune information retrieval applications, and not as a mechanism for invasion of privacy.

52 Many personal workstation operating systems can now be equipped with virus protection software which can detect and warn the user of unexpected *modifications* to **critical** files on the user's machine, but I have never seen one which monitors access. The user does have some countermeasures, such as keeping critical files on a separate disk and never mounting that disk while running software that he or she does *not* trust, or encrypting critical information when it is not being used, but the cumbersome nature of these measures makes them impractical outside of very high security environments with very security-conscious users.

Who Owns Access Histories?: Privacy and Market Research

The analysis of consumer behavior has become a major focus of attention in the business world. Supermarkets have on the one hand implemented laser scanners that track the products being purchased by shoppers (and linked them to systems that automatically issue a set of custom tailored discount coupons at the checkout register) and on the other hand now encourage payment with credit cards, allowing the development of databases that track consumer purchases in tremendous detail [Mayer, 1990]. Companies like American Express that have access to extensive histories of customer spending practices and preferences are now marketing finely tuned customer lists to their business partners-for example, I might receive mailings from American Express that offer me airline upgrades on airlines that I don't fly regularly, based on statistical analysis of my purchasing profile which indicates that I spend over $25,000 per year on airline tickets and that none of these charges go to certain airlines. Similarly, in many industries there is now an intense focus on what goods are selling, and in what marketplaces, and this information is employed in very complex pricing decisions (consider again airline seats as an example.) The practice of "data mining" from customer histories has begun to be viewed **as** simply effective exploitation of previously untapped corporate assets [Piatetsky-Shapiro & Frawley, 1991]. In addition, we are now seeing considerable use of multi-source data fusion: the matching and aggregation of credit, consumer, employment, medical and other data about individuals. I expect that we will recapitulate the development of these secondary markets in customer behavior histories for information seeking in the 1990s; we will also see information-seeking consumer histories integrated with a wide range of other sources of data on individual behavior.

The ability to accurately, cheaply and easily count the amount of use that an electronic information resource receives (file accesses, database queries, viewings of a document, etc.) coupled with the ability to frequently alter prices in a computer-based marketplace (particularly in acquire on demand systems that operate on small units of information such as journal articles or database records, but even, to a lesser extent, by renegotiating license agreements annually) may give rise to a number of radical changes. These potentials are threatening for all involved. Consider just a few examples:

- For the first time, libraries should be able to **easily** collect reliable data on how often specific journals are read, or even the pattern of access to specific articles within these journals. This information can be used to decide not to subscribe to journals, which worries the publishers.

- Publishers can employ this usage information to set prices on journals or even specific journal articles based on popularity. This leads to price instability, which worries the libraries.

- While citation data as a measure of the impact of a publication has been controversial (though it is already considered in tenure and promotion decisions at some institutions) usage data is less ambiguous; if nobody reads a publication, it is unlikely to have had much impact. This is of great concern to authors.

- Usage data makes it much easier for authors, publishers and libraries to rapidly reflect the short-term interests of the user community by keeping track of what is popular and trying to produce or obtain more of it.[53] To some extent, this is at odds with the development of the scholarly record and the integrity of scholarship. Archival publications are not necessarily read a great deal, but some would argue that it is of vital importance that they continue to exist.

- There is a tendency in systems that stress popularity to ultimately reduce diversity; if everybody else is reading something, then one concludes that one needs to read it also. The temptation to select as one's reading the ten most popular articles of the week is very dangerous to the development of a diverse body of ideas. It is also worth noting that producers of abstracting and indexing databases are increasingly considering the subscription patterns of libraries in deciding what journals to cover; this seems to make the databases more marketable. If these producers were to emphasize heavily read journals, these abstracting and indexing databases will tend to become less comprehensive guides to the literature (and, indeed, pathways to material in less well known journals).

- There is a danger that the system of statistics collection can be manipulated by those that understand it. This can range from authors repeatedly accessing their own works to get their statistics up through more sophisticated approaches (for example, including many popular keywords in an abstract even if they have little to do with the actual subject of the work so that many people will retrieve and view the work).

These examples have emphasized applications of data about the use of information resources such as viewing or downloading journal articles. However, the availability of searches is also of great value: it tells information product designers about the kinds of information that people are looking for, and also the means that they are using to locate it. This is invaluable market research data for designing new information products, and for marketing and improving existing ones.

The ability to collect not only information on what is being sought out or used but also who is doing the seeking or using is potentially very valuable information that could readily be resold, since it can be used both for market analysis (who is buying what) and also for directed marketing (people who fit a certain interest profile, as defined by their information access decisions, would likely also be interested in new product X or

---

53 It is interesting to note how each technological development that undermines privacy seems to be complemented by a technological countermeasure that supports privacy. Consider the case of pay telephones. At one time, these were a wonderful way to obtain anonymity; one simply deposited cash. and the source of the call was untraceable. Now, of course, most **pay** phone users are using credit cards because they are so much more convenient, not realizing that if they use these cards all their calls can be tracked in great detail. (In fact, many public phones will not even take cash anymore, due in part to the expense of collecting the cash and the fact that the cash is an invitation to vandalism.) In France, vendors now offer a phone card which has a specific "cash" value; one pays cash for it, and it is debited as one makes phone calls using it. This is a form of "electronic cash" which facilitates anonymity. (It also has some other important advantages; for example, while one **can** loose the card, one cannot incur the virtually unlimited bills against one's account that can be caused by a stolen phone credit card number.) Another example of this technological balance is the development of Caller ID facilities by the phone companies; these were quickly complemented by facilities that allowed a caller to block the display of the Caller ID to preserve anonymity.

special offer Y). While such usage (without the informed consent of the recipient of the advertising) may well offend strong advocates of privacy, in many cases the consumers are actually quite grateful to hear of new products that closely match their interests. And libraries and similar institutions, strapped for revenue, may have to recognize that usage data can be a valuable potential revenue source, no matter how unattractive they find collecting, repackaging and reselling this information.

Competitive intelligence is a burgeoning field promoted by any number of consultants. One aspect of competitive intelligence is knowing in what areas competing corporations (or, in academic world, research rivals) are seeking information. For example, it is valuable to know, if one is a coporation in the pharmaceutical industry, that a competing corporation is seeking articles about the effects of a given drug. Of course, once one recognizes that one may be a target of competitive intelligence, it is possible to deliberately offer disinformation that will lead the competition to an incorrect assessment of one's interest, and even deliberately send a competitor down false trails. Clearly, the type of information that can be collected about information seeking and use in the networked environment is invaluable for competitive intelligence. And it is worth noting that even fairly highly aggregated information can be of value in a competitive intelligence activity: for example, from the aggregated article access information for a given university (without any indication of who within that university accessed the material) it is quite reasonable to draw conclusions about the research directions of specific research groups that are very likely to be correct.

Some of these examples seem farfetched. But consider a number of trends. As electronic information providers license information rather than simply selling it, they can *require* usage reporting as a condition of license.[54] This is done in other areas. Libraries and universities are both aggressively seeking new ways to generate revenue; the resale of statistics about electronic collection use and/or searches, particularly if they can satisfy themselves that some level of privacy is being maintained by not including the identity of the user (if they know it) could be a very attractive revenue source. It is unclear whether these institutions have any legal obligation to ensure confidentiality of this information.[55] If one signs a contract with a commercial information service such as Dialog, issues of confidentiality can be negotiated in advance as part of the contract; but when one is accessing (anonymously or otherwise) a public-access information service, it is unclear what to expect, and in fact at present there is no way to even learn what the policy of the information service provider is. As the secondary markets develop it is even conceivable that when accessing a for-fee resource one might pay more for privacy, and that when accessing a public-access resource the user's client and the server for the public-access system might well negotiate various levels of confidentiality (no logging, statistical compilation only, actual

54 Such conditions could be imposed either directly on a library licensing the information for local mounting, or, less visibly, thorough contractual constraints on a third party such as Dialog or OCLC that makes the information available to the library community.

55 Resale of information about who is searching what type of information is not the only issue. A financial information service might be a good investment for a brokerage house or a merchant bank for example; they might internally exploit knowledge that could be gained from records of what customers were searching information about what companies or products.

text logged for searches but without ID, no resale or reuse outside of the internal operations of the information provider, etc.)

A final aspect that should be mentioned is that in the print world the library served as a shield for its user community in the sense that it purchased materials such as journal subscriptions. The act of purchasing and the cost of purchase might well be an act of public record discoverable under a Freedom of Information Act. But only the library knew who was using the material, and that information (in the form of circulation records) was protected. Further, because information was acquired in highly aggregated forms such as an annual subscription to a journal, the act of purchase revealed very little about the interests of the library patrons—indeed, there is no a priori reason to assume that purchasing decisions are always directly driven by the short term needs of specific patrons in the case of a research library. Now, consider the electronic world, where a library frequently acquires rather specific information (such as a single article from a journal) in response to the specific request of a user. This purchase, as an external business transaction, may be a matter of public record. Further, if the end user rather than the library as intermediary acquires the article, it may be possible to rather directly link information use to individuals or departments. The electronic information environment may well call for considerable reassessment of the definition of public records, particularly in the context of state universities, as these records are defined by federal and state Freedom of Information Acts.

The uses described for information about searches and access patterns are simply extensions of well established practices such as market demographic analysis and competitive intelligence. New uses for this information, unique to the networked information environment, are also being researched. For example, Professor Mike Schwartz at the University of Colorado has been exploring the concept of resource discovery-automated methods of discovering or locating potentially interesting network resources [Schwartz, 1989]. One of the techniques that he has studied is the examination of access pattern of other members of a user% affinity groups; for example, a botany professor might be interested in resources that other members of the botany faculty are utilizing regularly but that he or she is unfamiliar with. Such research may ultimately lead to new tools for locating information resources which will call into question the appropriate balance between privacy, competition, and cooperation in various communities.

Privacy, Intellectual Property and Electronic Mail Enabled Communication

Electronic mail based discussion groups—sometimes called electronic journals in academic circles if their editorial policies parallel those of traditional printed journals —have become extraordinarily popular on the Internet. These fall into two major categories— LISTSERVs and mail reflectors. Mail reflectors are simply special user IDs; when one sends electronic mail to such a user ID it is redistributed to the subscribers of the mail reflector automatically. Maintenance of the subscriber list is typically done manually or semi-automatically, with the convention being that if the mail reflector's address is of the form user@ hostname then there is an additional mailbox in the form user-request@ hostname to which requests to join or leave the mail reflector are directed. LISTSERVs are based on a program that was originally developed for the IBM

Conversational Monitor System (CMS) environment.[56] There are thousands of such mailing lists on the Internet; in addition, many such lists are reciprocally gatewayed to Usenet newsgroups, which are essentially very large collections of publicly readable electronic mail messages that are propagated through the Internet (and beyond), organized by topic (newsgroup name). Like electronic mail lists, some Usenet newsgroups are moderated; others are completely open. Unlike directly electronic mail enabled services (LISTSERVs, LISTSERV imitators, and mail reflectors), Usenet newsgroups do not appear to the reader as electronic mail, but rather as continually updated databases that are viewed through a program such as a newsgroup reader; electronic mail only comes into play when a reader wants to enter a Usenet Newsgroup discussion by posting a message. The privacy and intellectual property aspects of these mailing lists and newsgroups are very interesting, and probably largely ignored by most participants in them.

Some lists deal with topics that are controversial—for example, the Usenet news group ALT.SEX.BEASTIALITY or the recently established LISTSERV on gay and lesbian issues in librarianship, to mention only two examples.[57] Currently, USENET newsgroups offer a moderate degree of privacy; a newsreader running on a client machine uses a protocol called NNTP (Network News Transfer Protocol) to pull list postings down from a local NNTP server to the user's client. The list of newsgroups that the user is interested in is stored on the client, and one can find out what a given user is interested in only by looking at his or her preference files on that client machine or by monitoring data transfer from the NNTP server, both of which are relatively difficult.[58] LISTSERVs, on the other hand, require interested parties to actively subscribe in order to receive electronic mail that is posted and include options which allow anyone to look at who has subscribed to a given list (except for those users who have explicitly chosen to conceal their identities; these individuals are invisible except to the system administrators or list administrators). Of course, the vast majority of LISTSERV subscribers are blissfully unaware of the fact that their identities can be easily discovered, or of their option to conceal their identity. It is only a matter of time, in my

---

56 More recently, software has been developed for the UNIX environment which emulates most of the functions provided by the LISTSERV program in the IBM CMS environment.

57 There has been enormous controversy about the appropriateness of various universities making such newsgroups available to their communities; these have been well documented in the Computers and Freedom mailing list postings by Carl Kadie. To my mind, these controversies help to illustrate the gulf between the library tradition of not only intellectual freedom but of free access to information and the values of the computing establishment. It seems likely that if these were print works that were owned by the libraries of the Universities in question there would have been little debate about the right of these libraries to own them as part of their collection and to make them available to the university user community; this would have been a clear case of intellectual freedom on the part of libraries. But when such resources are made available thorough institutional computer systems (where there is little philosophical basis established for determining appropriate content) major controversies quickly erupt.

58 Interestingly, traffic analysis between client and NNTP server is probably easier than breaking into the file system on the client, and this can be viewed as another illustration of the way in which the deployment of the distributed computing environment has exposed individual's activities to much more scrutiny. If the client is on a large time shared host then it is not clear why specific newsgroups are being transferred to that timeshared host; if the client is on a personal workstation, however, it is relatively easy to *assign* responsibility for the transfer of material from a specific newsgroup. There is also a program called "arbitron" written by Brian Reid which publishes regular statistics about the usage levels of various newsgroups; this is again, I believe, based on traffic analysis techniques.

option, before some organization begins to make use of LISTSERV subscriber lists as a means of identifying groups of individuals that the organization wishes to communicate with. In a very real sense, one can view the LISTSERV system as a very public definition of the interests of many of the individuals on the network today. Put simply, one monitors a Usenet newsgroup, and the fact of that monitoring is between the user and the local Usenet distribution host; one subscribes to a LISTSERV and the fact of that subscription is generally known throughout the Internet, unless the subscriber takes a positive action to conceal it.

A second issue has to do with the ownership of material that participants post to these discussion lists or newsgroups. Currently this is a highly contentious issue, and positions range from organizations that sponsor discussion lists (such as the Well service in the San Francisco Bay area, which simply states that posters own their own words) through individuals who argue that they automatically own their own words and affix copyright notices to postings stating this option. When one considers the text of a LISTSERV discussion in the aggregate, it is even less clear who owns rights to complications copyrights. While a rather complex consensual etiquette is developing which suggests that one should not repost from one list to another or reuse a list posting without the author's permission, the legal (as opposed to consensual and moral) basis of these conventions remains extremely unclear. Many LISTSERV are beginning to view this in some sense as a contractual matter; upon subscription they present subscribers with the assumptions about reuse of postings on the list.

9. The Internet Tradition of "Free" Information: Quality, Integrity and Liability Issues

Fee based services are a relatively recent development on the Internet. Prior to the last few year, for both policy and historical reasons, such services did not exist on the net; certainly, there were machines, services and resources that were restricted to specific user communities (for example, super computer centers where time was allocated through a grant-like mechanism, or machines that belonged to specific universities and were used by communities at that university), but this was considered to be a very different situation from a vendor that provided service to anyone on the net who was willing to pay for such service. The recent presence of commercial information providers such as Dialog and BRS indicates that these policies are a thing of the past, and that current policy at the very least welcomes vendors supplying services to the research and education community. However, there is a strong philosophical bias towards the use of "free" information on the network among most of the network user community. This is a particularly comfortable fit with the values of the libraries that have been appearing on the network both as information providers and information organizers: the Internet tradition of free information is consistent with the library ethos of both intellectual freedom and free access to information. And there is a great deal of free information available; in fact, much of the development in software tools (which were themselves typically free, public domain software, at least in their initial stages) to facilitate the mounting of networked information resources (for example, WAIS and Gopher) has been to facilitate the ability of organizations on the network to offer access to an ever growing array of publicly-accessible networked information resources. This bias towards free information is evidenced by the rather minimal billing and access

control facilities in these software systems,[59] and indeed throughout the Internet generally.

While there are any number of organizations which have the dissemination of information to the public as part of their mission (including most types of government), it is important to recognize that the strong bias in the Internet user community to prefer free information sources provided by these organizations is not without problems. These problems include a tendency by network users to use relatively low quality information (a "you get what you pay for" argument), a lack of accountability for the quality and integrity of information offered without charge to the Internet community, and the potential for various forms of bias to find their way into the most commonly used information resources on the network. The ultimate result a few years hence—and *it may not be* a bad *or inappropriate response, given the reality of the situation-may* be a perception of the Internet and much of the information accessible through it as the "net of a million lies", following science fiction author Vernor Vinge's vision[6O] of an interstellar information network characterized by the continual release of information (which may or may not be true, and where the reader often has no means of telling whether the information is accurate) by a variety of organizations for obscure and sometimes evil reasons.

The first issue with "free" information is that it is, of course, not really free, but rather subsidized. Free information might be subsidized by a government agency as part of that agency's mission. It might be subsidized by a not-for-profit organization as part of that organization's mission to communicate its viewpoint to the public. A university might make information available as part of its missions to support research, education and public service. Some public information resources might be subsidized by a for-profit corporation as part of a public relations campaign.[61] It might, as discussed elsewhere in this paper, be provided as a means of acquiring market research data or mailing lists of people with specific interests. Following traditions in both the print and broadcast media, it might be subsidized by advertisers as a means of delivering advertising.[62] As an extreme case, one can envision the Internet analog of television

---

59 There is some evidence of a new focus on fee-for-service information resources on the network. The University of Minnesota, which funds Gopher development, has recently begun the implementation of licensing agreements for the Gopher software that assess substantial charges for organizations that wish to provide information—particularly for-fee information—using the Gopher software, in conjunction with an upgrade of the software to Gopher+, which includes facilities to address billing and user authentication. While Thinking Machines Corporation placed the initial version of WAIS in the public domain, Brewster Kahle, one of the original developers of the system, has recently formed a company called WAIS Incorporated which is seeking to commercialize the system (or at least the server software) and to work with information providers who wish to offer their information through WAIS servers-often for a fee.

60 This is described in his 1992 novel *A Fire Upon the Deep* [Vinge, 1992]

61 Print publications have tried to establish conventions that clearly identify advertising material as advertising; for example, when a corporation purchases space on a newspaper's editorial page for a corporate statement on a public-interest issue, the print publication typically goes to some lengths to try to indicate that the material is not part of the publications editorial content but rather paid "advertising" (communications). There will clearly be a need to develop similar conventions for Internet information resources.

62 Advertiser-supported material might be viewed with particular caution. A number of authors have explored the effect of advertising subsidies on the popular media (both print and broadcast) and have suggested that advertisers have a significant effect on content and editorial positions taken by these

"infomercials" where one obtains some information (probably of questionable accuracy and/or value) along with a very long sales pitch on some given topic, such as getting rich through selling electronic real estate to house out of copyright books. In a sense, one can regard much of the current crop of "shareware" and demo versions of commercial software as forms of advertising promotions.

It is also difficult to entirely separate "free" content from the mechanisms that provide access to the content. One of the properties of networked information distribution is the ability to suddenly and simultaneously make new information available to an enormous user community; a community that is perhaps far larger than the ability of the computing system supporting the information resource to service at initial peak load. New documents, new virus definitions for a virus protection program, new software releases or bug fixes may be provided free by the information provider, but the public access resources supporting access to this material may saturate under the demand levels of initial public release.[63] In these situations, users who have a real need for timely access to the newly available information may pay a premium to some access provider (perhaps a service like CompuServe or Applelink) rather than retrying and continually being refused access to some public FTP archive. Or they may be willing to accept some advertising, or the collection of their address for future marketing purposes as a condition of obtaining timely access to the information.

Another very real issue is lack of responsibility and accountability in making information available on the networks. Tools like WAIS and Gopher have made it very easy for anyone to offer an information resource to the network community at large; one simply implements a WAIS server or a Gopher server on one's personal workstation, for example, using publicly available software. Whether this information is accurate, and whether the institution or individual that initially made it available feels any responsibility to ensure that it is accurate or current is unclear. A recent problem that caused a considerable amount of discussion the LISTSERV PACS-L is indicative of the problem. Someone on the network went looking for a copy of the periodic table of the elements. Much to their delight, they located one that someone had made available through one of the networked information access tools. Unfortunately, upon closer inspection, this periodic table was missing entries for a number of the elements. Unfortunately, it was not clear that anyone felt much responsibility to remove the incorrect information from the network, or to update it to be accurate. While the readership of the list PACS-L now

---

media. See for example *The Media Monopoly* [Bagdikian, 1992]. One of the most pernicious aspects of this advertiser influence is that it is hard for most viewers to identify and subtle in nature. In a networked information environment where advertising may be more easily ignored by viewers of information resources, it may be even less clear who the advertisers are.

63 The accepted community practices for access to this type of information distribution are complex and quite interesting. For example, some public FTP archives limit the number of concurrent accesses from a given institutional network, with the idea being that if information on the archive is very heavily used by a given user community, that community should import the information and then make it available internally to reduce load on the public FTP archive. Unfortunately, there is little automated assistance to facilitate such actions; in an ideal world, an institutional network might recognize that a file was being frequently requested from a (globally accessible) public FTP archive and automatically import (cache) the relevant file and then redirect requests for copies of the file to the institutional file server; later, when demand died down, the local copy would be discarded, and requests would go to the globally available archive. But this type of automated implementation of institutional responsibility for sharing in the resource commitment to distribute such files does not exist today.

probably realizes that they should not trust the information in this public-access periodic table, it seems probable that any number of new citizens of the network will rediscover this inaccurate resource in the coming months and years; hopefully, they will quickly realize that it is unreliable, but there are no guarantees. This problem is closely related to a number of currently unclear issues having to do with liability for incorrect software or databases.[64] While liability for failures to provide correctly functioning software and accurate electronic databases are likely to provide fertile ground for litigation in future, many of the issues in this area are focused around some sort of implied quality or appropriateness of a product in the context of a sale or licensing agreement: it is particularly unclear how these issues apply to information resources that are provided on the "use at your own risk" basis that characterizes many (particularly not institutionally supported) "free" Internet information resources today. [65]

10. Access to Electronic Artifacts: The Problems of Versions, Presentation and Standards

The Nature of Electronic Works

When we consider the contents of digital libraries, we encounter a bewildering menagerie of different artifacts: interactive services, computer programs, text files, images, digital movies, networked hypertext information bases, real time sensor feeds, virtual realities. The taxonomy for these artifacts has yet to be firmly established; indeed, one of the greatest sources of difficulties in developing a consensus vision of the electronic libraries we hope to make available through the NREN. There is some basic conception of a "work", however—not necessarily in the legal sense of intellectual property law, but rather an electronic "object" or "document" (in the broad sense), as distinct from a service. Such works can be viewed, copied or referenced; they may be electronic texts, or multimedia compound documents. The are the analogs and generalizations of books, journal articles, sound recordings, photographs, movies or paintings.

The printed page or image can be apprehended with no more technology than the human eye; other works are recorded on storage artifacts such as videotape or phonograph records which require technology to convert them to a form that can be apprehended by the human senses. Yet these storage formats and technologies are fairly well standardized (at least in comparison to the digital world), and, at least to some level of approximation we believe that the content of such a work can be moved from one storage medium to another (say, from 35 mm film to videotape or from LP recording to CD audio disk), perhaps with some degradation of quality but without

---

64 There is reasonably well established liability for provision of incorrect information in commercial settings, such as credit bureaus; the situation for individuals or even institutions that make information available for free (with no promises of accuracy or maintenance) is much less clear. See, for example, [Samuelson 1993] for a further discussion of liability issues.

65 Resources that are institutionally supported are likely to be of better quality than those that are support by the actions of an individual. But it is also important to recognize that there are a number of resources that are offered by third parties, particularly information access providers such as Dialog, that also have major problems with completeness and accuracy.

loosing the “essence” of this content. This relatively high level of standardization is facilitated by the fact that methods of presentation are well established and fairly simple-one listens to sound and views movies. The experience is not interactive. But, when we consider the new electronic “works,” it is clear that they can only be apprehended by the human senses and the human brain interacting with a computer system that includes software and various input and output devices. The experience of these works is complex and interactive; a work can be viewed or experienced in many different ways. Further, other intuitive measures of a work are lost; for example, browsing a printed work gives the browser a sense of the amount of information that the work contains. It is unclear how to measure the amount of information that is contained in a multimedia database.

There is enormous variation among the capabilities of these computer mediators, and too few (or too many) standards. Material may be converted and transferred or displayed in many forms; these transformations can cause major changes in both the presentation and the intellectual content and thus threaten the integrity of the work. Worse, in many cases the content of the work is inextricably bound with the access methods used to view (and navigate within) the work. It is impossible to separate content from presentation of viewing. This inability to isolate intellectual content calls into question the long term and broad based accessibility of works.

Presentation and Content Standards

Standards are needed that permit intellectual content to be encoded into a work in a way that is independent of any specific access software; this is needed to ensure that material can be viewed from any appropriately configured platform, and ensures that the work can outlive specific hardware or software technologies. This is both a preservation issue-ensuring that the work will be available for the foreseeable future-and a portability issue, ensuring that the work will be available from multiple hardware platforms today or in the near future. Today, few useful standards exist, and, for various reasons, those that do are not widely adopted as vehicles for distributing electronic information. Worse: it is becoming increasingly clear that there is a serious lack of consensus as to where intellectual content stops and presentation begins,[66] and the extent to which information providers such as publishers are prepared to distribute content as opposed to specific presentations of content. This controversy has had a considerable impact on the development and acceptance of relevant standards, and most standards currently seeing broad use tend to encode presentation along with content rather than separating the two.

66 This ambiguity between form and content has very significant implications fOr copyright. A number of information providers seem to be taking the position that by simply reformatting or marking up out of copyright works they can return the specific representation of the work to copyright control. This is not necessarily a bad thing, as the availability of out of copyright works in marked up forms such as the encoding specified by the Text Encoding Initiative program adds substantial value, and the copyright protection protects the investment of the companies doing the markup, but it is a source of considerable uncertainty and confusion.

A great deal of electronic information today comes packaged with access mechanisms. In CD-ROM publishing, access software is almost always part of the disk.[67] In other environments one finds information packaged in formats such as Hypercard stacks for the Apple Macintosh.[68] Increasingly, we are seeing the use of programs that provide access to books and journals—Superbook from Bellcore, the RightPages system from ATT Bell Labs and similar systems. Such integration of retrieval apparatus with content raises serious concerns about the ability of libraries to provide access to this information in the long term and across multiple platforms. At the same time, other information providers, such as the Government Publishing Office (GPO) and the Bureau of the Census are creating other problems by issuing data, such as the Tiger files, without any access software, which raises serious questions about the ability of libraries to meet the objectives that are mandated by program such as the Federal Depository Library legislation.

Purchasers of electronic information such as universities and libraries are particularly eager to obtain information in presentation-independent formats such as SGML that allow the information to be reformatted for a wide range of display devices, and even easily processed by computer programs that perform various manipulations or indexing on the information. By contrast, many publishers have expressed a great deal of concern about the loss of control over the presentation of their works; they are worried about both quality and integrity questions. Thus, publishers are in many cases much more comfortable distributing information in formats like bitmapped images or PostScript files that allow reproduction of the information only in a specific presentation (that is, the information can only be replicated either in print or in a screen display with exactly the layout, fonts, and similar properties that were defined by the publisher when the information was created). Another factor in representation choice is that formats such as bitmaps, while they preserve the integrity of the print page, require a much more sophisticated support infrastructure of storage servers, networks and display devices than simple, unformatted text, and thus bar a large number of current network users from access to the material. SGML, while less demanding of storage capacity and network bandwidth, requires sophisticated viewing systems which today serve as a barrier to very broad access.

Even the transfer of files in presentation-specific formats is problematic. Standards for bit-mapped images are still immature; while the Internet Engineering Task Force recently came up with a basic standard for monochrome bitmapped images of pages, [Katz & Cohen, 1992] there is still no agreement on the broader structure that should

---

67 This explains, for example, the great difficulties that libraries and other organizations have encountered in networking CD-ROM databases. Most CD-ROM software is designed to run in the PC environment which does not currently integrate easily or well with commonly used wide area networking technology (i.e. TCP/IP and related protocols). Yet the content of the CD-ROM is so closely integrated with the access software that it is not feasible for purchasers to write their own access software that is more compatible with the network environments in use within the purchasing institutions, and, in fact, the information publishers regard data about how their information is formatted on the disk that would permit the writing of alternative access and presentation software to be a trade secret in many cases.

68 I am not aware of any software in common use that permits browsing of Hypercard stacks on other hardware platforms such as the PC, although there are of course programs that provide the ability to construct and browse similar databases on the PC.

be used for sets of pages[69]. Reproducing such bitmapped image files exactly is still tricky due to variations in the resolution of display and printing devices, which may require interpolation or decimation of the image. The difficulties of printing PostScript files created on one system and then transferred to another system are well known, and include not only header file incompatibilities and PostScript interpreter problems but also problems with fonts.[70]

SGML is frequently suggested as a good prospect for a content standard that can also incorporate some presentation information. However, SGML is really a standard which defines a language for defining document schemas (called Document Type Definitions, or DTDs) and for marking up documents according to a given DTD. While there are numerous attempts to use SGML as a basis for developing industry or application specific document markup standards (for example, the publishing industry has developed ANSI/NISO Z39.59, sometimes called the "electronic manuscript format" which is aimed at the transfer of manuscripts from author to publisher and for use by publishers during the editorial process; the Air Transport Association has developed DTDs for applications such as aircraft maintenance manuals; and the Text Encoding Initiative is developing standards for deep markup to support textual analysis processes of various scholarly disciplines in the humanities and social sciences), it is unclear to what extent these will be accepted.

Currently, most authoring systems do not support SGML (though there is some modest evidence that this is slowly improving). Most documents are either authored using word processors—Microsoft Word, WordPerfect and many others-or using various markup systems such as Troff and Tex. There are converters that move from one of these formats to another, but usually with some considerable loss of information for complex documents. Further, languages like Tex and Troff have many variants and enhancements.

Effectively, if one looks at how documents are being distributed on the network today, there is typically a canonical version of the document that provides as much content markup as possible-typically right out of the authoring system that created it. Users of the same authoring system can use this version. Then there are typically two derivative versions-one in pure ASCII text suitable for viewing on lowest-common-denominator terminals—and one in a PostScript or other print-ready format that (with a good deal of

---

69 The IETF standard allows multiple page images to be transferred as a group, but access to a specific page within the group is awkward. Some publishers are distributing collections of pages as sets of files, with one page per file, using the IETF format. Other groups are looking at higher-level structures that can be used to transfer groups of pages.

70 Fonts represent a particularly subtle problem. While a given font is not protected, certain representations of font families can be copyrighted, as I understand it. Today, in order to facilitate reproducibility, many PostScript files are shipped with their font definitions (which make the files huge), violating the copyrights of the fonts' rightsholders. There are a series of technologies being deployed at present, such as Adobe's SuperATM and Acrobat systems which permit documents to be shipped without fonts. The receiving workstation can use some parameter information to "regenerate" or interpolate a font that is supposed to be similar to the font actually used in the document, thus avoiding not only the copyright problem but also the very real practical problem of receiving a document that employs fonts one does not have on one's local machine. However, this very convenient substitution is performed at the expense of degrading the integrity of the document, and, again, most users may not really even be aware of what is happening, other than that the document looks a little strange.

luck) can be used to produce a printed version of the document. Most of the presentation, and often a great deal of the content, is lost in the ASCII document; the print-image document may or may not be usable, but if it is it preserves presentation integrity at the cost of making most of the intellectual content inaccessible.

While image information does not face the same dichotomy between content and representation a number of the same themes reappear. There is an enormous proliferation of formats for various types of digital images—TIFF (various dialects), GIF, PICT, etc.—which can be intermixed with various compression schemes (CCITT Group 3 or 4, JPEG, etc.). The formats themselves are in some sense more of a nuisance than anything else, and there is software available that converts from one format to another without much, if any loss of information. The compression standards are another matter entirely; JPEG, for example, includes "lossy" compression modes in which the size of the compressed image can be traded off against accuracy of reproduction, with more compact images offering a less detailed reproduction of the original image. Accuracy of reproduction of images—particularly color images-depends of course on having a sufficiently high quality display and/or printing device. But there are also more subtle issues: different monitors, for example, display the color palette differently, and a painting will look quite different when displayed (at identical resolution) from one monitor to another. This is a substantial concern for many color imaging applications, particularly in the fine arts.

There are on the order of twenty different standards for the storage of digital sound at varying degrees of quality and compression; size of the stored sound object depends on sampling rate, dynamic range and the compression scheme used (and whether that scheme is lossy or lossless), among other factors. Many of these schemes are platform specific.[71] More general multimedia standards—for example, for video or other compound objects—are still in their infancy, with a number of platform-specific standards such as Quicktime for the Apple line and Video for Windows coming into use.[72]

As one views these standards in the context of digital library collections one can see a conflict between integrity and access emerging. While it seems likely that over the next few years the industry will continue to adopt more platform-independent standards and that at least some publishers will move towards standards for electronic text that provide more content information and are less presentation specific, it is improbable that at any time in the near future we will find that the user community has installed a base of access devices with homogeneous, high-quality capabilities for reproducing

71 Part of the problem here is that hardware for sound reproduction is not universally available on various platforms (it is now standard within some vendors' product lines, but not across vendors) and the hardware used to play back sound is also not well standardized yet.

72 Quicktime is now available on Windows platforms, at least for viewing movies, and it seems likely that over time most of what come to be the better established digital movie standards will be ported across multiple platforms. This will become easier as the processing to support material that is formatted according to these standards is more readily done entirely in software with acceptable performance and reproduction quality. Right now, these formats push the capabilities of the basic hardware on most common platforms, and often benefit greatly from specialized hardware assistance. It also remains to be seen whether translators from one format to another become readily available for movies as they have for the popular image formats.

sound, displaying color images or viewing digital movies. It is also important to recognize that the capabilities of a user workstation are not the only issue in full quality reproduction of electronic information; network bandwidth is also a factor. The larger the object the longer it takes to transfer it. A high quality digital image or sound recording will take longer to move to the user's workstation for viewing than a lower quality one. In some cases this may translate into a cost issue; if the user is paying connect time for some service, or paying for network bandwidth on a usage-sensitive basis, he or she may be unwilling to pay for a full quality image, or may wish to browse lower-quality simulacra before selecting objects to move in full resolution. For some services, bandwidth limitations may be absolute in the sense that using a full-quality service will make it too slow to be usable. Consider users connecting from home; in most cases they are limited to rather slow connections (9600 bits/second in most cases today, even with relatively expensive and modern modems; perhaps in the not too distant future ISDN at 64 or 128 KBits/second will become a generally available and affordable reality); at these speeds, many users find it more satisfactory to communicate with systems by emulating older character-based terminals rather than running available graphical interfaces based on the X Window system technology. The issue is not that they cannot support the X Window system on their workstations; rather, it is that they don't have the bandwidth on their network connections to run it effectively. It seems likely that libraries, as institutions, will obtain high-bandwidth links to the Internet much more quickly than most end users. and that the issue of available bandwidth to the end user will be a critical factor in determining when information can be delivered directly to the end user in electronic form; it is likely that there will be a long transitional period during which libraries will have to act as staging sites for information on its way to the end user, or to some printing facility in cases where the end user does not want to move the information all the way to his or her workstation prior to printing it.

Thus, information servers such as libraries will face the need to decide, as they design their systems, how much downward conversion they are willing to do in order to permit viewing of material in a degraded format, and how they will make the user of that material aware that he or she is indeed receiving a degraded presentation. Some degraded versions of content will obviate the need for specialized software viewers, thus adding value. Also, while the libraries or other information providers may make the decisions about the spectrum of capabilities available in access systems, rights holders—publishers and authors—will clearly, through license agreements, have a voice in the extent to which libraries will be allowed to apply these capabilities to deliver degraded-quality representations of works to users who cannot view the "canonical" full quality version of the work.

The Problems of Structured Information

There is a natural tendency to view the contents of digital libraries as primarily information intended for direct human apprehension—text, sound, movies and the like. In fact, it is clear that a major class of electronic information resources will consist of structured databases—not databases of abstracting and indexing records or of fulltext, but of genetic sequences, weather observations, commodities prices, chemical compounds and their properties, menus, plane schedules, biographical summaries, and thousands of other groups of information. This information may be viewed by humans

(with the aid of a viewing program) but also be used extensively by other programs that manipulate information, summarize it, reason with it, and correlate it with information from other sources on behalf of human users. In order to be able to access and operate on such information reliably it needs to be rigorously structured; the human eye and the human mind are far more capable of compensating for variations, applying common sense, and inferring than most software. There are a growing array of tools available for defining data structures that can be moved across the network, such as Abstract Syntax Notation One (ASN.1); however, there will be a growing need to standardize semantics for data elements that may appear in records retrieved from information servers on the network. Indeed, the success of such standardization efforts will be central to the success of large scale, distributed scientific programs in areas as diverse as systemic biology and botany, molecular biology, global climate change, and the initiative to map the brain. While some communities, such as the library community and more recently the various groups interested in electronic data interchange (EDI) have substantial experience in this area and have made some progress within limited domains, this is a new concept for many scientific communities and one that these communities are having considerable difficulty making rapid progress in realizing. Further, while many of these scientific and research communities clearly recognize the need to develop such standards for data elements and a clear focus exists within the research community to address the issue, it is far less clear who will do the work to develop and maintain similar standards for the more mundane or commercially-oriented standards .[73]

Sadly, until such time as these structured information interchange standards become defined and widely implemented, we will be unable to realize one of the major potential payoffs of creating libraries of electronic information—the ability to view these as knowledge bases that can provide a foundation for the development of more capable software agents that can act on our behalves, but that can function relatively independently, without the need for a human being's eyes, mind and judgment to preprocess large amounts of information.

11. Digital Images and the Integrity of the Visual Record

In a very important recent book titled The *Reconfigured Eye: Visual Truth in the Post-Photographic Era* Mitchell, 1992], William Mitchell traces the historical role of the visual image as a record of events. Prior to the development of photographic technology, painting served the cultural role of recording things, events and people. In the 19th century photographic methods were developed which were far more accurate than painting in reproducing and recording; indeed, the author argues that the development of these technologies freed painting to explore new and less literal reproductions of visual reality. While "trick photography" and various photographic effects were developed very early in the evolution of photography, there seemed to be a fairly well understood implicit consensus that these uses of photography to capture unreal images

[73] Indeed, for the commercial standards, in the worst case we are likely to see a number of different commercial information suppliers attempting to establish incompatible de facto standards in order to capture markets and ward off competition. The notion that open standards actually establish and foster the growth of markets in information is not yet generally accepted by most of the commercial information suppliers.

would be applied primarily in artistic frameworks where their use was clearly identified to the viewer of the photograph. In the 20th century, as photographs (and later moving images) became an increasingly central part of advertising, alterations to photographs without warning to the viewer became more commonplace.[74] In the past few years, however, photography has begun to migrate to digital technology; this shift, in conjunction with the ease with which digital images can be altered, combined, or even generated from scratch by computer graphics programs has created a situation where the implicit evidentiary value of images (without regard to their provenance) has been lost.

It is not entirely clear that this loss of evidentiary value has fully penetrated the public perception, despite recent developments such as advertising which interposes modern-day celebrities into historical films in commercials, very sophisticated special effects in movies, and even virtual reality technologies to which the public is now regularly exposed. But the fact remains that it is almost impossible to tell whether an image represents a record of a real event given current technology; moreover, while twenty years ago it required a great deal of sophistication to alter images or films, easy to use software that can be employed to perform such image manipulation (and, increasingly, image creation) is now widely available on inexpensive personal computers.

This change in the meaning of images has a number of implications for digital libraries. When viewing an image, one must always harbor a certain degree of skepticism about whether the image actually represents a real event or thing and to what extent this representation may have been altered. In essence, when using images, one must constantly be concerned with the source of the image, whether the image has been altered subsequent to its capture, and the purposes for which the creator of the image is making it available. Without facilities to track and verify the source of images, they have become meaningless as a part of the historical record. 75

12. Authenticating versions and sources

In the print environment a great deal is taken for granted about the integrity of documents. If an article appears in a journal it is extremely rare that the authorship that the journal lists for the article is called into question; when this happens it is typically framed either in the context of scientific and scholarly misconduct such as plagiarism and/or results in a lawsuit. Outside of scholarly circles the issue is probably more likely to revolve around the publisher's right to publish (that is, whether the rights holder indeed gave permission, or whether the claims to rights on the material are valid—for example, in the case of unpublished archival material) than whether the author

[74] While advertising was perhaps the greatest culprit in the undermining Of the Integrity of visual images as a record of events, there were many other contexts in which altered photographs were used: politically motivated changes and sensationalistic news reporting are two other common areas.

[75] Of course, just because an image cannot be verified as a representation of an actual thing or event does not mean that the image is not of interest. Computer generated or computer altered images are of vital importance as works of art, as hypothetical models of things (for example, a reconstruction of a dinosaur) and as records of culture. The problem is that the viewer of the image needs to know the context within which to understand the image.

attribution is false. With the exception perhaps of certain tabloids one would not normally assume that there was much reason to question authorship, Similarly, print publication naturally tends to produce specific editions of a work; if two people both have the same book or issue of a journal there would be little reason to question whether the two copies of the publication were indeed identical in content.[76] It is not that a publisher couldn't deliberately publish variant copies of what is labeled as the same edition, or deliberately misattribute authorship of material, but rather that it does not happen often and when it does the publisher is typically readily identifiable and can be sued by the aggrieved parties. Further, there is little motivation (other than general malice) to motivate most publishers to do this; a publisher would have to go to considerable trouble, expense and risk in order to do it.

Perceptions and concerns in the world of networked information are quite different. It is very easy for someone to distribute information over someone else's name, and hard to trace the person who does it in most cases. It is very easy to replace an electronic dataset with an updated copy, and, since there is no automatic system which distributes multiple copies of the original version to different sites (such as the distribution of an issue of a printed journal or a book) the replacement can have wide-reaching effects. The processes of authorship, which often involve a series of drafts that are circulated to various people, produce different versions which in an electronic environment can easily go into broad circulation; if each draft is not carefully labeled and dated it is difficult to tell which draft one is looking at, or whether one has the "final" version of a work. Because of the ease with which material can be taken from one document and inserted into another which can then be circulated to a large number of people quickly, there are concerns about quotation from obsolete or draft ("unpublished") versions of a work. Visionary authors such as the late Ithiel De Sola **pool**[77] have written that the world of networked information would lead to the **demise** of the "canonical" form of documents that print publication created, and that documents would become living texts that were continually adapted and annotated. Events thus far have suggested that De Sola Pool may have overstated the case. While it is common within small groups to have people annotating drafts of a document, they are typically ultimately brought to a final, "canonical" form. Further (and it is unclear whether this is due to current limitations in information technology such as groupware and collaborative authoring systems or whether it is a more basic problem having to do with the limits to the size of a group that can collaborate effectively) such continual annotation typically occurs among a relatively small community of collaborators or reviewers, and not among the full community of interested individuals on the net. In cases where a large scale, long term collaborative effort is taking place to develop and manage a "living document" such as On Mendelian Inhetitance *In Man* at Johns Hopkins, a fairly complex and formal editorial structure is set up to manage this collaboration, and considerable care is taken to validate and track updates and their

76 While there would be little question in most people's minds about what they could expect from printed literature, this is not necessarily the case with other media, such as videotapes, audio recordings, or computer programs, although even here it would probably not occur to most people to doubt that identically labeled copies of a work were in fact identical.

77 See his work on "The Culture of **Electronic** Print", reprinted/adapted in his book *Technologies* of freedom [de Sola Pool, 1983].

sources. In a real sense, these efforts are as much undertakings to create and manage databases as they are efforts to author documents.

Interestingly, aside from a few pranks and malicious acts (most commonly people sending electronic mail with false origin addresses) it is unclear whether the fears that people seem to harbor about the deceptive and mutable nature of the electronic environment are justified by real occurrences of problems. Also, those publishers who have risked the Internet environment have had less problems with piracy than one expect given the experience of software vendors, for example. The simple fact is that several people are successfully distributing publications on the Internet today for profit (though I don't know how much real profit they are making, they are still in business after several years in some cases). Nonetheless, it seems clear that if network based information distribution is to become a widely accepted context for the sorts of archival materials that libraries currently acquire and provide access to in print these concerns must be addressed. Certainly, the development and wide implementation of technologies and practices to address these concerns will, at the least, lead to a far more robust and reliable environment, although from a strict cost-benefit analysis it may be difficult to fully justify the costs of addressing some of the fears one hears voiced.

To clarify the focus here, it is important to recognize that while the network will undoubtedly be used extensively for transacting commerce (including, as just one example, commerce in the form of acquisition of electronic information by individuals and organizations, which may involve activities such as the identification of the parties involved, the exchange of credit information for billing purposes, the assessment of charges against some type of account, and even the acceptance of the terms of a license agreement for copyrighted material) and there is, I believe, strong justification for ensuring that these commercial transactions are conducted in a technical environment that protects the security and confidentiality of all parties, the issues involved in protecting transactions are somewhat different from those involved in ensuring that a user of the network who finds a document somewhere can verify that the authorship attribution is true and that the copy which the user is looking has the same content as the version that the author "published." The issues of protecting network commerce generally are outside the scope of this paper, other than simply to observe that for a market in network-based digital information to develop it will be necessary to develop and implement adequate measures to protect commerce on the network and also to conduct some form of electronic rights clearance. The remainder of this section will address issues of verifying authorship and integrity of contents, and the state of the art in technologies to accomplish these objectives.

Public key crptography and various higher level protocols layered above the basic cryptographic algorithms offer methods that can be used to effectively address both of these needs. A public key cryptosystem can be used to attach a signature to a digital object in such a way that the contents can be associated with a given individual or organization at a given time. There are well-established algorithms for computing "digests" of digital objects in such a way that it is extremely unlikely that any change to the object can be made without changing the value of the digest computed from it. Thus, by checking whether the digest for an object in a user's possession is the same as the digest value that the author has signed and makes available as the signature of the current version of the work, it is straightforward to check whether one has the same object as the author or publisher distributed. These systems offer the additional feature

of non-repudiation; it is possible to include capabilities so that one can prove that a given author actually distributed a document at a given time even if that author later denies it. Such capabilities can be seen today in the Internet in the privacy-enhanced mail system [Balenson, 1993; Kaliski, 1993; Kent, 1993; Linn, 1993].

While the basic technology exists to solve the problems in question (at least as long as one is satisfied with literal bit-by-bit equivalence of two digital objects as a definition of having the "same" document, which is often really overly restrictive, since it prevents any reformatting, character code conversion or other activities that might be needed to successfully move the document from one machine to another, even if these do not change the "content" of the document in any way) the operational problems of implementing these technologies on a large scale in environments such as the Internet are far from solved. There are least four barriers:

- Standards are needed. While the algorithms are well understood at a general level, in order to ensure interoperability among implementations agreements need to be reached at a much more specific level of detail and documented in standards. Parameters such as the precise types of signatures need to be defined, along with lengths of public/private key pairs, the exact computational algorithms to be used, and the supporting protocols and data interchange formats. The IETF specifications for message digest algorithms [Kaliski, 1992; Rivest, 1992a; Rivest, 1992b] are an important step in this direction, but more work is needed. It is also important to recognize that there is more to an effective system for authentication and verification than simply algorithms; there are application protocols which use these algorithms to be defined, along with an accompanying infrastructure of service providers (see below). An additional problem that must be resolved in the standards area is the seemingly continual conflict between the standards proposed or established by the Federal Government through the National Institute for Standards and Technology (NIST) and the standards that are favored by the commercial and research and education communities .[78]

- Patent issues need to be addressed. What is currently widely accepted as the best public key cryptosystem is called RSA (named after its inventors, Rivest, Shamir and Adelman); this was patented and commercial rights to this patent are licensed, as I understand it, to a company called RSA Data Systems incorporated. Similarly, Public Key Partners holds patents to a variety of public key and message digest algorithms; to make matters even more confusing the National Institute of Standards and Technology (NIST) has filed patent applications for some of the algorithms that it has developed and adopted as federal standards and proposed licensing these (on an exclusive basis) to Public Key Partners; again, there seems to be a provision for free use of the patents for at least some types of personal or non commercial use. The net effect of these patents on basic cryptographic technology is to promote considerable uncertainty about the status of the algorithms and to inhibit their incorporation in software of all types (but particularly public domain software; while large corporations can negotiate and pay for

[78] Recent examples of this problem include NIST's controversial adoption of the Digital Signature Standard Algorithm and the Secure Hash Standard (FIPS 180), as well as the widely-denounced proposal for the Clipper chip and its accompanying key escrow system. The continued unwillingness of the government to recognize the RSA public key algorithm as a standard despite its widespread use is another example.

license agreements with the rightsholders, individual software developers or university based software development groups that wish to distribute their work without charge typically are unable or unwilling to do so), despite the relatively liberal positions that the commercial rightsholders seem to be taking on personal use and use by the research and education communities. Further, the patent filings by NIST are regarded with considerable suspicion in some quarters; there is concern that in future these patents might be used as a means of controlling the use or implementation of the technology.

- Cryptographic technology is export restricted. This has caused two problems. The commercial information technology vendors have been somewhat reluctant to develop products which can only be marketed in the United States without major export complexities, particularly given that authentication and digital signature technology tend to be very basic building blocks for distributed systems. In addition, because the world of the Internet and of networked information is very clearly viewed as a global rather than a national enterprise, system developers and standards makers have been reluctant to use technologies that cannot be freely used internationally. The issue of the justification and implications of applying export controls to cryptographic technology is a very complex one that is well outside the scope of this paper; however, the impact of the current restrictions must be recognized. In addition, it should be observed that while the position of the United States on the export of cryptographic technology is crucial because of the nation's leadership in so many areas of information technology, other nations may also have laws related to the import, export and use of cryptographic technology that also create barriers to the free use of authentication and digital signatures on a global, Internet-wide basis.[79]

- Critical mass and infrastructure. Like so many things in the networked environment, these technologies will not come into wide use unless they are available on a large part of the installed base. Authors want to communicate; publishers and libraries want to make information available. If this information is not readily used without specialized cryptographic software that is difficult and/or costly to obtain, or that cannot be used outside the United States, it is unlikely that authors, publishers or libraries will use them. While in the specific applications under discussion here of verifying authorship and integrity of objects it should not be necessary to have specialized cryptographic software support simply to view the material but rather only to conduct verification,[80] it is really more of a question of whether the investment is worthwhile because enough

[79] There is a major public policy debate currently taking place about the appropriate balance between the rights of individuals and the private sector generally to privacy on the one hand and the desires of law enforcement and intelligence agencies to be able to monitor communications on the other. While the details of this debate are outside the scope of this paper, the interested reader might wish to review the history of export restrictions on the RSA algorithms, the recent proposal by the Clinton administration for the Clipper chip, and the deployment of the PGP ("pretty good privacy") computer software both inside and outside the US.

[80] In cases where cryptographic technology is being used in conjunction with rights clearing some proposals do call for the distribution of encrypted documents that cannot be read without both software to decrypt and the appropriate key. Somewhat similar approaches are being used today where vendors will distribute a variety of locked digital information on a CD-ROM (such as programs or font libraries) and then issue decryption keys on a file-by-file basis as the customer purchases these keys; one advantage to this approach is that one can phone order the information by providing a credit card and getting a key, without waiting for physical delivery of media containing the information being purchased.

people will make use of the services. In addition, it should be recognized that there is a substantial infrastructure needed to make public key cryptosystems and the applications constructed using them work on a large scale, including key providers, certification authorities, directories of public keys, and "notary public" servers (third parties that can witness signatures and contracts, or that can record the fact that a given entity had a given document at a given time and date as a registry function). The precise details of this infrastructure will vary depending on the standards, protocols and procedures that develop in support of an implementation of the technology; how these details change from one proposal to another are not important here, but recognizing that an investment in support infrastructure must be made is vital. Further, as indicated earlier under the discussion of standards, the conflicts between federally endorsed standards and standards favored by much of the user community are having the effect of fragmenting and confusing the user community, and greatly delaying the achievement of the necessary critical mass.

It is interesting to place the issues of authenticating authorship and document integrity in the broader context of the way in which migration to a networked information environment is beginning to suggest an "unbundling" of the services that publishers have traditionally provided in the print world. Print publishers serve as selectors of material; they prepare the material for distribution, distribute it, manage rights and sometimes royalty payments, and authenticate authorship and integrity, among other functions. In the network environment, it is clear that distribution (at least through a mechanism as crude as making a document available for anonymous FTP) can be done by anyone. It is clear that services that help people to identify material of interest such as abstracting and indexing services, reviewers, bibliographers, and ratings servces are likely to play an enlarged role in the electronic environment, and that these services can be quite separate from the persons or institutions that make material available. It may well be that authenticating and verifying the integrity of a document is at least optionally a separate, and separately priced (and perhaps costly!) service from simply obtaining a copy of the document.[81] If so, it will be interesting to see how much use is made of such services (outside of some specific environments, such as litigation, where cost is typically not much of a factor and such issues simply must be unambiguously established) and particularly the extent to which people are willing to pay to allay their fears about the electronic information environment.[82]

81 The integrity of published works is not an entirely new problem. A number of publishers currently provide loose-leaf services for areas such as tax law; determining whether a user has the most current version of such a service is today an important question with a potentially high payoff.

82 It is also necessary to consider other implications of establishing a chain of provenance for an electronic document. As discussed previously, technology for tracing the provenance of a document is well established, and depends on sophisticated cryptographic technology. It seems likely that US government, and perhaps other governments have agencies that are monitoring most international traffic, and that any encrypted traffic will attract their attention., since at least at present it is relatively rare In some nations use of cryptographic technology may be illegal, either across international boundaries or even within the national boundaries. Even if it's not illegal, it may attract attention. Are scholars prepared to attract the attention of such communications security agencies as a consequence of maintaining verifiable ties to the scholarly record?

13. Citing, Identifying and Describing Networked Information Resources

As networked information resources are integrated into the body of information, both scholarly and popular, it will be necessary to extent traditional print-based methods of citation to accommodate reference to these new network-based resources. Here, the objective is to continue the functions served by citation in print literature: to permit a work to make reference to the contents of another work with sufficient specificity to permit the reader to obtain a copy of the cited work and locate the part of that work being referenced; to give the reader of the citation enough information to make some judgments about whether he or she is already familiar with the cited work, and to provide some information about the cited work such as date of publication, title and author which might help the reader to determine if it is worth obtaining a copy of this cited work. It is important to note that traditional print citations today serve both of these purposes; for example, citations consisting simply of document numbers assigned by some document registry are not typically used because while they would allow the reader of the citation to obtain a copy of the cited document, they don't tell the reader anything about the cited work to help in making a decision whether to obtain a copy of it. 83

At the same time that the need to cite electronic information resources is being recognized, several other closely related requirements are emerging. These include the desire of libraries, bibliographers and other organizations and individuals that organize information to catalog the increasingly valuable and common electronic information resources; essentially, to extend the existing mechanisms of bibliographic description and control to facilitate access to these resources. The needs here are closely related to those of citations, but more extensive in that there is usually a requirement to include more information about how to obtain access to a given resource once identified, and also requirements to include subject access or other classification information.

It is interesting to note that both for citation and cataloging purposes a number of people have expressed a desire to have the citation or cataloging record include some information (such as document digests or signatures, as discussed earlier in this paper) that would permit the user to check that he or she had retrieved the same version of the electronic object that the creator of the citation or descriptive record had originally described (at least as an option: when one is talking about electronic documents this makes sense, but when one is making reference to a database that is continuously updated at the level of an information resource, rather than referring to the contents of a specific record in that database at a specific point in time, such version information does not make sense). Logically, this requirement makes little sense. Reasoning by analogy with the print world, if a citation specifies the second edition of a specific work, it is possible that the publisher might change the contents of the work and reprint it without updating the bibliographic information or date of publication, in effect creating two editions of the work that have different content but are not identified as distinct

83 It is worth noting that in some areas of scholarship historically citation systems have been used that only address the identification of a work, or passages from it, without referring to specific editions. Examples include biblical scholarship and some types of literary criticism. Usually in these situations there is an implied canonical text, so it is not necessary to specify the specific edition of the intellectual content.

editions.[8] However, this does not happen often (at least for materials that are extensively cited and where very precise citation is important) in the print world and people don't generally worry about it much.[85] The emergence of this requirement for version verification in the electronic information world simply underscores the general perception that electronic information is more volatile and more easily changed, and that the contents of electronic objects cannot be trusted to retain their integrity over time without introducing special verification processes into the system of access and management of these resources. It is also worth recognizing that on a technical level this problem of version verification is largely unsolved as yet; while the digital signature and digest algorithms discussed earlier can readily ensure that a document is bit-for-bit the same as the one cited, citation typically is more concerned with intellectual content. As we move to an environment where software and protocols for retrieval of electronic documents (in the broad sense of multimedia objects) becomes more adaptive and mature, transfer of documents from one host to another may commonly invoke format translations and reformatting of various types automatically,[86] while such translations would presevee the intellectual content of the document (perhaps at varying levels of precision, depending on whether the transformations were lossless and invertable), the transformation would of course change the actual bits comprising the document and thus cause it to fail a version comparison test based on such bit-level algorithms.

---

84 T. be clear: current library cataloging rules direct catalogers to explicitly differentiate works that are different even if the publisher has not done so.

85 Indeed if anything, the problems today with citation to printed material, as discussed earlier in this report include the difficulty that the creator of the citation often does not realize that the publisher is producing multiple editions targeted for different geographical regions or for different subsets of the readership (for example, trade magazines that include special advertising sections targeted at readers who work in specific industries) and hence doesn't create a sufficiently specific citation. From the publisher's point of view, there is often great economic incentive to keep repackaging and reissuing content with minimal changes as new editions or even new works; the notion of going to the trouble of producing an unadvertised and unlabeled new edition and quietly introducing it into the marketplace is relatively rare, at least for print; this practice does occur sometimes with electronic publications such as software, where minor corrections or improvements are sometimes shipped automatically without much publicity, although even there the publisher usually changes the version number. There are a few examples of audio materials where different versions have been shipped with the same cover and same publisher catalog number.

Also, in the print world, in cases where a citation is to a work where there is some question about the precise final form of the work, conventions have been developed such as indicating "unpublished draft" or "in press" to alert the user of the citation that there may be some problems. Of course, such citations are the exception rather than the rule.

86 T. provide only a few examples of such translation, a document might be changed from one character set to another (ASCII to EBCDIC or UNICODE); fonts might be substituted, since fonts are copyrighted and the workstation receiving a document might not have the fonts used by the author, so it might be necessary to substitute similar fonts that are either in the public domain or that are licensed to the receiving workstation; an image or digital sound clip might be converted from one format to another, and the resolution or sampling rate might be altered; or more extensive format changes might occur, such as the rendering of a postscript document into a bitmapped image prior to transfer. The extent to which these transformations preserve the intellectual content of the work are highly dependent on the nature of the transformation and also the use to which the document will be put when it is transferred; for example, if it is only to be viewed, then a transformation from SGML markup to a bitmapped image makes no difference to the content in some sense, but if that same document is to be edited or analyzed by a postprocessing program, then there is a very large loss of information in the conversion from SGML to bitmapped representation.

A third set of requirements are more technical in nature but address some of the needs for both cataloging and citing networked information resources; while they solve neither problem they provide tools for developing solutions. In addition, a solution to these technical requirements is needed to enable the widespread development and deployment of a number of important networked information applications. These technical requirements are based on the need for standards so that one object on the network can contain a computer-interpretable “pointer” or link to another object on the network. This is needed for network-based hypertext systems such as the World Wide Web. It is needed so that document browsers can automatically follow references in a document when these references are to other network resources. It is needed so that bibliographic or abstracting and indexing records that describe electronic information resources can include information about where to find and how to access these resources. This last case is particularly important for a number of projects that are now underway where large bitmapped image databases of material are being created; because of the size of these databases it is desirable to store them at only at most a few sites on the network and to retrieve page images from them on demand; yet multiple databases of descriptive records, developed by multiple organizations, need to include links to these image databases. Further, in some cases, the descriptive records are being distributed under different license terms than the actual content; for example, some major publishers are exploring scenarios where they give away brief records analogous to tables of contents in printed journals, and then charge transactionally for retrieval of the actual articles.

The idea is that these pointers to networked information resources should be representable as an ASCII text string, permitting their inclusion in both electronic documents and in printed documents, as well as their easy transfer from machine to machine and from one application to another within a machine (for example, via cut-and-paste facilities now available in most graphical user interfaces, with the idea being that a user might view a document or an electronic mail message or a screen display from an online bibliographic database in one window, find a reference to a document that he or she desires to fetch, and simply highlight and drag the citation to another application, which would then fetch the object or open a connection to the service, using whatever access protocol is required).

These technical requirements are being addressed by a working group of the Internet Engineering Task Force. While the technical details of the IETF standards proposals are beyond the scope of this paper (and, indeed, some specifics of the standards are still under active debate within the IETF Working Group **as** of this writing) there seems to be some substantial consensus on the overall approach to be taken. It should also be recognized that there are some very substantial research problems in dealing with these technical requirements in full generality, and thus the IETF work should be regarded as a beginning and a framework that will undoubtedly undergo a great deal of extension and refinement in the coming years based on operational experience with the first generation standards, improved understanding of the theoretical issues and abstract modeling questions underlying the standard, and the continued development of protocols and applications for accessing networked information resources of various types.

Roughly, the IETF proposals call for the definition of a syntax for what they call a locator, which is an ASCII string that identifies an object or service that is hosted on a specific machine (typically specified by its domain name) on the network, the service (such as FTP, electronic mail, Z39.50 database query) that is used to obtain the object, and the parameters that are to be passed to that service to identify the specific object to be obtained (for example, in the case of FTP the fully-qualified filename). There are several problems with locators as a basis for citation, however. Machines on the network come and go over time, and files are migrated from one machine to another. Some commonly used files are duplicated on multiple machines; from the point of view of citation, one wants to refer to content and not instances of content, and thus should no more list machines containing copies of a file than one would list libraries holding copies of a book in a citation. An object may be accessible through multiple access methods (for example, FTP and database retrieval); indeed, the method of access may change over time and in response to improved technology, but the content being accessed remains unchanged. Further, one cannot tell whether two different locators actually refer to the identical content.

Thus, the IETF working group has proposed the definition of identifiers, which are strings assigned by identifying *authorities* to refer to content. An identifier for an object, then, is just a two-component object consisting of a specifier for the identifying authority (these would be assigned centrally, as a service to the Internet community, much like top-level domains or network numbers) and the identifier that the authority provided. These identifying authorities (and other organizations) may offer services that provide a mapping from an identifier to a series of locators, which could then be used to actually obtain access to a copy of the object. Some mapping services, particularly those operated by specific identifying authorities, might only resolve identifiers assigned by the operating identifying authority; others, perhaps operated by organizations such as libraries, might attempt to resolve identifiers issued by multiple identifying authorities into sets of locators. Locators would be viewed as relatively transient; at any time one could obtain a fresh set of locators corresponding to an identifier. Identifiers would be used in citations and other applications. It is important to note that the IETF model explicitly recognizes that deciding whether two instances of an object are “identical” is a subjective issue which is highly dependent on the objectives of a given identifying authority, and that there will be a multiplicity of such identifying authorities, which might include publishers, service organizations, libraries, or industry-wide standards implementations (such as the International Standard Book Number in the print world). The same content might be assigned identifiers by multiple identifying authorities; in some cases two objects might be viewed as identical by one identifying authority (meaning that the authority would return locators for both objects in response to its identifier) and yet viewed as distinct by another identifying authority. 87

Rules for citations are typically set by editors of journals, or sometimes by professional societies (for groups of journals) or by style manuals (such as the *Chicago Manual of Sty/e).* While a number of journals (both print journals and electronic journals) have

87 AS a specific case in point, one identifying authority might view a bitmapped image and a Postscript file of the same document as identical; another might view these as different objects. The issue of format variations and the extent to which these variations, as well as multiple versions of documents, should be recognized by and integrated into the locator and identifier scheme is still an active area of discussion.

already defined practices for citing electronic information resources it seems very likely that these practices will be altered over time to include identifiers which followed the IETF standard in order to facilitate both the identification and the retrieval of electronic objects. As these identifiers come into wide use, some of the other material that is currently specified in citations to electronic resources (such as the name of a machine holding a file available for anonymous FTP and the file name) might well be dropped. Some of the traditional citation data elements that help the reader to identify and evaluate the intellectual content of the cited work, such as author, title, and publication date, will almost certainly be retained. A few data elements used in some citation formats, such as the number of pages in a work, are problematic in an electronic environment; while it is clearly useful for the reader to have some sense of the size of a cited work, it is unclear how to most usefully measure this in an electronic environment that may contain multimedia works. The transformation of citation rules is likely to be a gradual process; it is important to note that, at least in practice, citation formats are really not national and international standards, but rather working rules that serve various communities, and there are a fairly large number of citation formats in common use.

Cataloging practices for networked information resources is an area that is currently under very active discussion. Several groups within the American Library Association (in particular, MARBI and CC: DA) are studying this issue and working on guidelines in association with groups that include the Library of Congress, OCLC, the Coalition for Networked Information, and the IETF. Some of the issues involved here are very complex, and not yet well understood; indeed, some of the questions involve very basic considerations about the purposes and objectives of cataloging. Taxonomies for classifying networked information resources are also needed, and still poorly understood. The current drafts [Library of Congress, 1991b; Library of Congress, 1993] from the American Library Association's MARBI committee again recognizes the use of the IETF locator and identifier structure as an appropriate means of encoding some needed information, and foresees a conversion to these standards as they are established, while also supplying some provisional field definitions that can be used by catalogers who wish to experiment with cataloging network resources in the interim.

It is also important to recognize that cataloging is only a part of the broader question of how to provide information to help users to identify and select networked information resources. Cataloging is concerned primarily with description and organization of materials (for example, through assignment of subject headings within some classification structure and vocabulary, or through the development of name authority files that bring together works published by the same author under different names, or different variations of a single name); equally important information which would allow someone to obtain evaluative information about a resource or to compare one resource to another is outside the scope of cataloging. Such information is provided by book reviews, consumer information services, ratings services, critical bibliographies, awards given by various groups, sales figures and other tools. All of these services-and new ones, such a certification that software works properly in a given environment or is free from viruses, for example—will need to be evolved into the networked information environment but with some new and challenging additions. One key objective will be to preserve, and if possible to expand the diversity of evaluative sources that information seekers can consult if they wish; just as one promise of the networked information

environment is an increased pluralism in available information resources, a parallel diversity of facilities for selecting from these information riches is essential.

The tools and methods of selection and evaluation must become more diverse and flexible. Today, virtually all evaluative information is intended for direct human consumption; a person reads a review or rating service and then perhaps makes a decision to acquire a product or use a service. It seems clear that in order to manage the overwhelming and dynamic flood of information that will occur in the networked environment we will need to develop software tools to help us in selecting information resources and navigating among them. Encoding and knowledge representation for evaluative information, and in fact even the definition of appropriate data on which to base selection decisions are areas in which research and innovation are desperately needed, along with all of the accompanying issues of algorithm design for software to assist in such decision making ; indeed, the lack of progress in this area may prove to be a significant limiting factor achieving the promise of a large scale networked information environment.

14. Directories and Catalogs of Networked Information Resources

As networked information resources multiply, one of the central issues will be locating appropriate resources to meet various needs for information [Lynch & Preston, 1992]. There are many tools that have evolved for identifying various types of information resources for various purposes, and many organizations that produce these tools for many reasons.

Libraries have played a role in this area through their collections (and the choices they have made in selecting and acquiring these collections), their catalogs, and the bibliographies and directories that they make available to their patrons. However, in the electronic environment, the role and content of these tools for locating and identifying information are changing. One important and problematic issue is the relationship between library catalogs and networked information resources. In the print world, one can distinguish the cafalog, which describes and provides access to material held by a given library from the *bibliography,* which defines and provides access to the literature on a given subject without regard to where that literature is held (and typically does not provide the user of the bibliography with any information that would help this user in physically obtaining access to material listed in the bibliography) [Buckland, 1988] .88

88 Basically for economic reasons, the coverage **of** library catalogs is typically limited. Since the early part of the century, libraries have typically been unable to afford to catalog the individual articles in journals that they receive, so they only catalog at the journal level. Bibliographies (or abstracting and indexing databases, which are simply the electronic successors to printed bibliographies) are used to obtain access to journals at the article level; library **catalogs are then used to determine if the library holds the journal containing the desired articles. So-called online library catalogs today typically at large research libraries offer access not only to the library's catalog, but also to some abstracting and indexing databases (bibliographies); a few systems offer the ability to view the bibliography as a form of catalog by permitting users to limit searches to articles in journals held by the library.** This is accomplished by having the library's online information system link the library's catalog database to the journal titles covered by the abstracting and indexing database. A few systems, such as the University of California's MELVYL system, or OCLC's EPIC/FirstSearch service have gone a step further and also linked the journal holdings of other universities to these bibliographies, thus in some sense transforming the bibliography into a union catalog of holdings in a specific discipline (though not a comprehensive one, since there are undoubtedly journals

Some leaders in the library community have discussed the transition to networked electronic information as a transition from the role of libraries in creating physical collections to a new role as providers of access to information that may be physically stored anywhere but is available through the network. In this new environment, the role of the library catalog in permitting users to identify relevant electronic information is problematic. One scenario calls for libraries to include in their local catalogs descriptions of networked information resources that the library chooses to logically "acquire" (either simply by selecting them and placing descriptive records for them in the local catalog, or in the case of fee-based services paying some type of license fee, or subsidizing transactional usage fees on behalf of the library's user community in addition to adding the descriptive record to the local catalog). An alternative scenario calls for libraries to simply provide their users with access to external catalogs, directories or bibliographies of networked information resources and to assist patrons in accessing these resources; in this scenario the "selection" or "acquisition" decisions of the library are accomplished at two levels: first, by the choice of external databases that they offer their patrons which describe available networked information resources, and secondly by the extent to which the library allocates both staff and financial resources to helping patrons to use different networked information resources, and to subsidize the costs incurred by use of these resources. Complicating the picture in either case is the inevitable development of various directories and bibliographies of networked information resources by other organizations that will be accessible to the library's patrons, in some cases for free and in other cases for fee.

It is also important to recognize that there will be a lengthy transitional period where libraries may provide access to directories of information resources and abstracting and indexing databases in electronic form, but during which most of the primary material, such as journal articles, will continue to exist in printed form. Linkages from electronic directories, bibliographies, abstracting and indexing databases, and online catalogs to the print holdings of libraries will be of central importance for at least the next decade. Experience has shown that these linkages are difficult to establish without human editorial intervention by simply matching on unique numbers such as the International Standard Serials Number (ISSN); yet the establishment of such linkages reliably will be of central importance in providing access to current library resources. Additionally, such links are essential in making effective, economic interlibrary loan and document supply services feasible.

Realistically, it seems likely that libraries will seek a compromise solution with regard to the representation of networked information resources in their local catalogs, probably including descriptive records for resources that they believe are important enough to spend money acquiring access to on behalf of their user community and for some carefully selected free public-access resources deemed to be of significance to their patrons. For access to other resources, patrons will be guided to external databases on the network, and libraries will develop policies about the extent to which they will subsidize and assist use of these external directories and the resources listed in them by various segments of the library's user community (in much the same sense that university research libraries today will go to considerable lengths to obtain access to

---

relevant to the discipline that are not covered by the producers of the abstracting and indexing databases). So, there is already growing ambiguity as to the boundaries between bibliographies and catalogs.

arbitrary material through interlibrary loan or purchase for faculty,[89] for example, but might charge students for a similar service if they offer it at all).

Not all identification or use of networked resources will take place through libraries, of course. Just as today people also identify and/or acquire material by reading advertising, browsing in bookstores, scanning book reviews, joining book clubs or by word of mouth, similar routes will be taken to electronic information resources. The only cause for concerns here are those of balance. While university research libraries are actively addressing access to networked information resources, the vast majority of public and school libraries lag far behind and lack the resources or expertise to address these new information sources; indeed many such libraries are today struggling just to survive and to continue to provide their traditional services. For many people without access to major research libraries, the primary routes to identifying networked information of interest may not be through libraries at all, but rather through information services on the network. But the level of these network information services has been disappointing, up till now; perhaps in future competing commercial services will improve the level of service, but at the cost of reducing equality of access.

But consider: while libraries, depending on their mission, budget, and patron community will vary in scope and depth of collections, one of the primary tenets of library collection development is to provide a broad, diverse, and representative selection of sources on areas that are within the scope of the library's mission. It is unclear to what extent other groups providing directories of networked information resources will reflect these goals of libraries; some directories will undoubtedly be forms of advertising, where a resource provider pays to be listed and is listed only upon payment of such a fee. Some databases of resources may be essentially the electronic analog of bookstore inventories, with all of the criteria for inclusion that such a role implies. Other directories may be built as “public services” by organizations with specific agendas and specific points of view to communicate. Services will develop that provide very biased and specific selection criteria for the material that they list in their directories; this will be a very real added value for their users, who in some cases will pay substantial sums for the filtering provided by these review and evaluation services There is nothing wrong with such directories; indeed they provide real value, offer essential services, and also ensure the basic rights of individuals and organizations to make their points of view

89 Specific mention should be made of the changing nature of the use Of interlibrary loan to Permit a library to obtain material on behalf of its users. Consider first the major research library; historically, interlibrary loan was used primarily as a means of providing fairly esoteric research materials to faculty when they were not held by the local library. With the growing inability of even research libraries to acquire a the bulk of the scholarly publications in a given area, we are seeing the use of ILL even to support requests from undergraduates. ILL is no longer used simply for esoteric research materials. Another important issue is the independent scholar—this might be an individual conducting independent research, a staff member at a small start up company that does not have a library, an inventor, or even a bright high school student: in all of these cases, the information seeker will most likely use a local public library and the ILL system to obtain access to the research literature. Such requests are relatively rare, and a decade or two ago were accommodated fairly routinely through the ILL system when they occurred; today, with the increased emphasis on cost recovery as a reaction to the overloading of the ILL system, the barriers to access by such disenfranchised patron communities are multiplying rapidly. There is a real danger than within the next few years the research literature will be essentially inaccessible to those library patrons who are not part of the primary user community of a major research library. This is a major threat to equitable access to knowledge, and one that may have some serious long-term societal implications, ranging from frustrating bright young students through handicapping the independent inventor or scholar.

known. But there is, I believe, cause for concern if the at least relatively "neutral" service offered by libraries is not among the options for seekers of information in the networked environment.

15. Conclusions

Integrity and Access Issues in the Broader Networked Information Context

Before attempting to summarize or draw conclusions from the material covered in this paper, it is vital to put the issues reviewed here in perspective. This report has concentrated on problems and open issues. In some cases it sketches a rather bleak picture, particularly in regard to the role of libraries as publishers move towards electronic information products. It has outlined a growing array of threats to information consumer privacy in the networked environment. Indeed, the purpose of the report is to highlight these issues and problems.

It is important to recognize and address these issues precisely because the potential of networked information is so significant. Realizing this promise is of central importance. Information technology and network-based access to a rich array of information resources can change our educational institutions (in the broadest sense not only of elementary and higher education, but of lifelong learning), our political system, our economic frameworks, and our culture. Visions of futures in which our children, anywhere in America, can browse storehouses of knowledge and cultural history available from electronic library collections, define goals which we collectively believe worthy; the question before us is how to achieve these goals. If the potentials were not so great, the issues defined here could be left to the evolving marketplace in electronic information and the continual redefinition of institutional roles that this marketplace is driving. But I believe that the promise of networked information demands conscious, deliberate choices, and, where necessary, investments to support these choices.

The other point that should be stressed is that we are in a very complex transitional period which is likely to continue to at least the end of the century. This is not only a transition from the traditional print publishing system (including the role of libraries in that system) to a system of electronic information distribution, but also to some extent a transition away from the existing system to new models for creating and controlling access to content. For example, government (at the federal, state or local level) may well commission the creation of content for use by the public, or license access to content on behalf of the public because access to this content is an essential element in the educational system (again, in the broadest sense of elementary, higher, and adult education). Authors may choose to make their creations widely available at little or no cost simply because they believe that access to these creations is of great importance to society, or because they are writing to communicate ideas rather than to make money. A new information distribution system, enabled by the ability of the network to make every participant a publisher and to disseminate materials in electronic formats widely and at very low cost, is starting to grow up alongside the traditional publication system even as this system of publication is itself transfigured. Depending on an author's goals in creating a given work, he or she may choose the traditional, copyright-controlled system based on publishers or the one of the new network-based

publication models as a distribution channel. Within this new, parallel, information distribution system using the network libraries will take on new roles and missions. This is a time of great creativity and experimentation, of exploring new roles and new models.

We are seeing signs that economics alone will not define the shape of the future. For example, in a networked environment there is a very strong tendency to centralize resources; the extreme case of this is the vision of a centralized electronic library in a given discipline that provides service worlwide.[90] While there are strong economic justifications for this sort of centralization in a networked environment since the presence of the network eliminates geographic-based use community affiliation and permits economies of scale that are amortized across national or international user communities, the predicted centralization is not clearly taking place. Rather, the networked environment is giving rise to a very pluralistic model of information storage and access; at one level, this is inefficient, as a good deal of information is stored redundantly, but at another level this is a comforting development since it re-enforces the value that we as a society place on distributed, democratic access mechanisms that lack central points of control. We have yet to fully comprehend the resolution of the conflicts between economics and cultural/institutional values.

Similarly, the destruction of the existing interlibrary loan system is not an entirely forgone conclusion; as authors, particularly authors of scholarly works, become more aware of the consequences of their actions, they are beginning to protest the confines of the existing scholarly publication system and in at least a few cases to explore alternatives, such as various forms of network-based electronic distribution of their works. There is a growing recognition that the publication system that has developed to support scholarship, teaching and research over the past centuries exists to seine these communities rather than to define their function. There is a perception within the research and higher education communities that they can define the future that they wish to live in, and that the members of these communities are responsible for defining that future. For example, I expect that there will be serious and occasionally bitter debates among the boards of scholarly societies in the next few years as the communities to which these societies are ultimately accountable wrestle with questions about whether these societies will have roles similar to for-profit publishers (perhaps subsidizing other activities of the society with profits from publication programs) or whether they will return to their original functions of facilitating communication and diffusion of new knowledge within scholarly communities, even if this means distributing their publications at little or no cost on the network and loosing the revenue that these publications generate (and presumably finding new financial models for supporting the society's activities and publications). This reevaluation of the roles of the existing system of publication in meeting the needs of the scholarly community is likely to be painful and acrimonious, since whatever their origins both commercial scholarly publishers and many professional societies which function as publishers are now very large and profitable businesses that will resist changes diminishing their size, income and influence.

---

90 Technically, such a facility is likely to be mounted on multiple hosts, probably at multiple sites, 'n order to provide some redundancy in case of disaster and to permit scaling to very large user communities. But, organizationally, the model is one of a single monolithic institution providing access to information.

As we look beyond the research and higher education communities, the picture becomes less clear, as the motivations of key stakeholders become more clearly profit-oriented and the sense of accountability to a community becomes weaker. When one considers the role of advertising, and the corporations that advertising serves in the development of the electronic mass media to date, one cannot be sanguine about predicting a future in which these media are held directly accountable for furthering the public good. Perhaps we can see the start of a divergence here between the research and education community and the general information consuming public (recognizing of course that many individuals participate in both communities to a greater or lesser extent at various points in their lives). The research and education community, which ultimately creates and can control most of the information it uses, is beginning to take responsibility for its own transformation into the networked information environment. On the other hand, the populace as a whole (including the public library system that serves this general populace) does not in any real sense create the information that it consumes, or control this information except in the most indirect ways (the power of the consumer's dollars in the marketplace and the power of the consumer's vote in developing public policy); content and the means of access to information are controlled by relatively unaccountable organizations like commercial corporations. In the general case, we are a society of information consumers who view ourselves at the mercy of information providers. The electronic information world of the general public may well be defined primarily by entertainment video libraries, interactive games, shop-at-home services that substitute for the printed catalogs that clog our mailboxes today, and "infotainment" segments advertising the latest in personal growth, weight loss, business success, and the like, with market researchers lurking in the wings to accumulate (electronic) mailing lists of qualified prospects. Here it is important that libraries, government information, and information from the scholarly community, as well as many diverse viewpoints from the general public on issues of importance maintain a presence among the information sources offered to the general public through the network, even if, following the patterns of today's broadcast mass media and print publications, such materials are only modestly used by the general public. Ensuring this continued presence is an important public policy objective. There is considerable precedent for this; for example, in the broadcast media the offerings of the Public Broadcasting System are not typically the highest rated programming, but they are offerings that make important contributions to our society in many different ways.

There is no question in my mind but that we will solve the problems and address the issues raised in this paper. The progress of information technology is inexorable; the promises and advantages compelling and the payoff enormous. It is clear that the private sector has now recognized the potential marketplace that networked information of various types represents, and has begun to commit massive financial resources to develop this marketplace. If not already the case, the scope of this private sector commitment will soon overwhelm the resources that the research and education community and the government have already contributed to seed and nurture development of the networked information environment. This will create additional pressures to address and resolve the issues quickly. It may also introduce a new pragmatism and expediency into the development of these solutions; while academics and policy makers sometimes debate issues at great length, the need to ship products,

launch services and recover investments is a great motivation to come up with some sort of practical solution and get it implemented in a timely fashion. The growing private sector pressures will also create considerable tensions and controversies, since solutions acceptable in the commercial marketplaces (and desired by the private sector) may not be entirely acceptable to the research and education community or to makers of public policy.

The challenge before us, then, is to ensure that we address the issues and solve the problems in the most timely way possible while, to the maximum extent possible, incorporating and balancing the interests and concerns of public policy, of the research and education community, and the private sector in these solutions. Speed is important; without timely progress we face the risk of being overrun by marketplace developments, which are not likely to reflect the balance of interests that I believe is essential for a future that will offer not only the commercial payoff but also the improvements in research, education, and the extent to which the public is informed. And balance is also vital: the interests of the various sectors involved are in many cases conflicting, and a deliberately and thoughtfully crafted balance among them will be needed to achieve the future that we desire. The importance of developing this balance is too great to be left entirely to the chance and marketplace forces.

Ensuring Access to Information in the Networked Information Environment

Publication, whether in print or in electronic form, is the act of making a unit of information available to the public, perhaps at some price. These individual units represent intellectual property for which the authors and/or publishers are frequently compensated. This is as it should be. At the same time, when all of these publications are aggregated, they form a major part of our societal, cultural and scholarly record and serve as a repository for our collective knowledge. Ensuring that our children, scholars, researchers, indeed all of our citizens, have some reasonable level of access to this collective body of information both when it first appears and even many decades later is a vitally important public policy objective. Today, this public policy goal is implemented by the provisions of the copyright law and by institutions such as libraries. The copyright law and the doctrine of first sale help to ensure that libraries exist and can effectively function; however, with some relatively modest exceptions, while the operation of libraries seems to be generally accepted as a public policy goal, the libraries of America are enabled more than they are mandated by specific federal legislation.

As this paper has shown, the mechanics of “publication”, its legal framework and perhaps even its definition are changing in important ways in the electronic environment. Further, as has been discussed, new forms of information rather different than the traditional published works collected by libraries are taking on increased importance: these include the contents of the electronic mass media and also the so-called secondary information sources (such as abstracting and indexing databases) which, when joined with the searching capabilities of computers, provide new and powerful tools for managing and navigating the growing primary literature. The public policy goals of creating and maintaining a reasonable level of citizen access to the published literature remain, but we may need to find new ways to achieve these goals. There are questions both of access to relatively current material and continued access

to the societal and scholarly record in the long term. This changing legal framework is making it very difficult for libraries to continue to fulfill the functions that they have traditionally performed in support of these public policy goals. Either changes must be made to permit libraries to continue to perform these functions, or some new or redefined set of institutions must be established and empowered to do so. There are many possibilities, some of which have been at least superficially mentioned in this paper, including changes to the copyright law (such as mandatory licensing), the creation of increased amounts of information or licensing of information at a national level, or changes to the depository provisions of the copyright law to ensure that copies of electronic works are registered with some institution responsible for their long term preservation. One can imagine a number of other legislative or regulatory approaches to addressing these issues.

One of the problems today is the general uncertainty surrounding intellectual property law as it relates to electronic information. This sense of uncertainty is both inhibiting progress and driving some developments that may well be undesirable from a public policy point of view (such as the increased use of contract law and licensing to control electronic information). Resolving these intellectual property questions in the courts will be a very slow and costly process and one that only increases the sense of uncertainty and risk surrounding electronic information. One alternative would be legislative action to clarify the issues and in some cases perhaps implement specific changes in support of public policy objectives. But, in an area as complex as intellectual property law changes will have to be made with great care and great wisdom; further, because intellectual property laws potentially impact so many areas of the economy and society (and also have important international implications) it may be difficult to develop a successful consensus on changes driven by the needs and public policy objectives related to networked information within this much broader community. There are other possible ways to make progress and reduce uncertainty, such as guidelines developed among the stakeholders which do not have the force of law, but which provide generally agreed upon rules of acceptable behavior; the model of the National Committee on New Technological Uses of Copyrighted Works (CONTU) with regard to the development of guidelines for interpreting the copyright law in the context of new technologies may be relevant here. CONTU both helped to clarify and obtain some consensus on issues, and also paved the way for subsequent legislative changes.

The purpose of this paper, however, is to make the reader aware of the growing problems in achieving the public policy goals related to access in an environment that is increasingly moving towards electronic information, and to provide background for an informed discussion of solutions, rather than to explore the ramifications of the various proposed solutions in detail. These problems are real and growing. But, I believe that our strength here is that as a society we have a reasonable consensus on the public policy goals, though there will always be debates about how much access is enough and how such access should be financed, as well as the nature of the implementation mechanisms and the continual tuning of the balance between rightsholders and the public.

Finally, I would note that federal government information has a very special role in the developing networked information environment. If it is made publicly available at little or no cost it will be a very widely used and important information source in the networked environment. Indeed, the creation and distribution of inexpensive high quality

information resources can be an effective instrument of public policy; one need only consider the enormous impacts that databases like MEDLINE from the National Library of Medicine and ERIC from the Department of Education have had in vastly improving access to and use of published information in the biomedical and health sciences and education respectively [U.S. Congress, 1990]. Federal leadership in information policy related to the electronic distribution of public information would also be helpful to state and local government in developing policies and recognizing the advantages that networked information access and distribution offer. Finally, large amounts of federal information can be used as a testbed for developing and proving standards, technologies and systems without the complexities, costs or limited and closed user communities that would typically be required if licensed commercial information was used in such experiments.

Privacy, Confidentiality and Anonymity in Access to Electronic Information

If there is relatively good consensus on the importance of access to information to our society, I believe that there is much less consensus about issues related to privacy, confidentiality, and rights to anonymous access. This lack of consensus goes far beyond simply access to the published works and to the societal and scholarly record, and is clearly seen in the many public policy debates related to privacy and confidentiality generally (for example, credit reporting, medical records, public records, computer matching of various types of files, debates about cryptography) as well as the conflicts between the cultures and perhaps values of libraries, the computing and computer networking communities, and the commercial world that have been illustrated here. The ability of information technology to provide easy access to and permit the analysis of vast amounts of information has implications that we are just beginning to understand. Further, as this paper has illustrated, there are many subtle questions related to the use, compilation, and analysis of histories of access to and use of information even in cases where users may be anonymous.

Hopefully, the paper will give the reader a sense of the scope, complexity and subtlety of the issues in this area. While perhaps there are a few areas, such as confidentiality of some types of records, on which there is general consensus and which might be addressed quickly, my sense is that it will be necessary to conduct an extensive policy debate with the objective of defining public policy goals before a great deal of progress can be made. In the meantime, to some extent, the best that can be hoped for is that users of electronic information become more aware of the privacy and confidentiality issues involved in their use of electronic information resources so that they can make more informed choices.

Many of the privacy and confidentiality issues discussed in this paper are peculiar in that they can be addressed on two levels: the legislative/policy level and the technological level. The technological solutions are often in turn driven by marketplace perceptions about the value of privacy and confidentiality; if consumers recognize that a serious problem exists and are sufficiently concerned to pay for the implementation of a solution, that solution will often become available. Legislation can, of course, also mandate the implementation of technological solutions, but this is rare. In my view, the technological solutions are often more robust than the legal ones, because the legal

restrictions are very difficult to enforce. Consider as an example the controversy about scanners for cellular telephones, and let us ignore the issues about consumer use of encryption and exportability of products incorporating cryptographic technology touched on elsewhere in this paper. A cellular telephone user concerned about privacy could purchase an encryption device which would provide a high assurance of privacy, at the cost of some inconvenience and subject to the limitation that secure communication would be possible only with other owners of a compatible encryption device. A few cellular phone users did so. Legislation was passed making scanners to eavesdrop on cellular phones illegal; however, such scanners were widely available and it seems likely that anyone who really wants one could still purchase or build one. So, the effect of the legislation has been to provide most cellular phone users (who have not purchased encryption devices) a false sense of security; while cellular phone eavesdropping as a consumer "sport" has no doubt been curtailed, I would suggest that the real problem hasn't been solved. A more effective solution would have been to either establish standards for cellular phone encryption and encourage the marketplace to implement them (and mount a campaign to make sure that users were aware of the risks of purchasing a phone that did not implement encryption) or perhaps even to mandate the inclusion of such encryption devices in new cellular phones.

Very similar problems apply in the case of services that require users to import software onto their personal machines for access, and where that software may collect and export information back to the service in question. While it might be possible to craft legislation to prohibit such practices, this would have to be done very carefully so as not to prevent legitimate and valuable applications. Further, as discussed, the consumer might well be willing to permit export of certain information in return for other considerations such as free or discounted access to services. A better choice here, in my view, would be to combine efforts to inform consumers (including perhaps some sort of labeling disclosure requirement on commercial software that exports information) with investment to develop good technology to permit the consumer to monitor and control the export of information from his or her personal machines. The difference between the cellular telephone example and many of the problems discussed in this paper, however, is that we do not currently have good technological solutions ready to deploy, and thus research investments are likely to be required.

Infrastructure and Standards

There is a tremendous amount that needs to be done to establish a viable infrastructure for electronic information and to ensure that it can become an effective, manageable part of our scholarly and societal record. A good deal of this work is neutral with regard to the public policy questions raised in this paper (although accomplishing these tasks will require that other public policy questions be addressed, such as those related to cryptography). Much of what is needed is simply funding (for research, experimentation and analysis and evaluation of experiments), standards development (discussed in more detail below), authoring and distribution of public domain computer software to help establish a critical mass of implementations in support of selected standards,[91] and to seed the construction of at least some parts of

91 There are numerous success stories in this area that deserve consideration. Software authored by universities and publicly distributed over the network without cost has led to the deployment of a number of important new network-based information services, such as Gopher. The availability of such software has

the infrastructure that will support networked information; the research and education community and the private sector are already working actively in these areas and are making considerable progress, as Section 2 suggests, but funding sources are few and funding is often a problem. I believe there is reason to be optimistic that some of the legislation currently under consideration will help to address these areas. Leadership in forging partnerships among the research and higher education communities, industry and government is also an important part of the effort required.

A few specific points should be emphasized with regard to the needs for funding. First, funding the infrastructure of the computer-communications network is certainly a prerequisite for the development of networked information, but there is additional infrastructure investment needed over and above that for the web of transmission facilities, switches and other technology necessary to create the communications network itself. This paper has discussed some of the areas in which investment will be needed, such as: systems to support integrity and authentication; systems to permit the location and identification of networked information resources; directories and catalogs to permit network users to find relevant information resources; systems to create, disseminate and view multimedia electronic works. Thus far, the vast majority of the funding invested in encouraging the development of networks at the federal level has gone towards building the communications infrastructure, and, while this investment has been quite successful to date (to the extent that there is serious discussion about when what parts of the communications infrastructure should transition entirely to the private sector) the facilities to support networked information are not nearly as extensive or advanced. The need for federal investment in the networked information infrastructure has not passed, and this should not be overlooked in discussions focused on the need for future federal support for the communications infrastructure.

Also, as a community I do not believe we yet understand how to solve a number of the technical and management problems related to networked information. There is a very real need for funding to support research and experimentation, including the implementation, testing and evaluation of a number of fairly large scale prototypes. The ability to test and learn from multiple approaches will be very important in guiding the development of technology in this area. In addition, we must be sure that there is funding not only for implementation but also for the follow-up evaluations and studies that permit us to really gain the full benefit from pilot projects. Further, there is relatively little basic theory to guide engineering projects in networked information, and much of the research in the field has a very pragmatic, near term focus on developing operational prototypes. A case can be made that this needs to be balanced by funding for more “basic” longer-term research.

---

served as a stimulus for additional software development by other institutions as well as widespread implementation of the services themselves. Industry has also made good use of this approach; one notable contribution here is the Wide Area Information Server (WAIS) system developed by Thinking Machines, Apple Computer, Dow Jones and KPMG. Finally, it is important to recognize that many people in the computer networking community believe that the funding that the Defense Advanced Projects Research Agency (DARPA) provided for the incorporation of the TCP/IP protocols into the UNIX operating system at the University of California at Berkeley during the 1980s was a critical factor in the success and explosive growth of both the Internet and the UNIX system.

Finally, this paper has not really discussed where the people will come from who will build and manage the networked information environment; while this is somewhat out of scope for a study of the integrity and access issues in electronic information (other than to point out the obvious, that there will be a need for trained and skilled individuals to manage the information and insure its integrity, and to help information seekers to gain access to it). From the point of view of developing the necessary technology and standards base and actually building the infrastructure, however, there is a developing shortage of people with the necessary combination of expertise. It is necessary for the higher education community to begin now to design and implement appropriate academic programs to develop a large pool of people who can contribute to designing and building the networked information enterprise; some universities have already begun this process, typically building on programs in library and information studies as a starting point. My view is that this is really in some sense a new field, though one that builds extensively on computer science, traditional library and information studies, communications technology, public policy and other disciplines. Funding to support academic research and the development of academic programs to support networked information can thus be viewed as part of the infrastructure investment that will be needed.

There are two particular problem areas impeding the development of the necessary infrastructure. The first, which has been discussed extensively in this paper, is the set of barriers surrounding the large-scale use of cryptographic techniques to implement the authentication and integrity functions that will be essential to the use of electronic information. The impact of these barriers is not limited to electronic information access and integrity; it also poses problems for a number of other network-based applications., including commercial transactions of various kinds. Resolving these problems, I fear, will require nothing less than the development of a rational, clearly articulated national policy on cryptography. There is, in my view, an urgent need for action in this area.

The second problem area is standards. As the paper has illustrated, standards are a key to developing the infrastructure, and also a central part of the strategy for ensuring that electronic information continues to be available in the fact of continual changes and improvements in the technology base used to house and deliver it. Yet the necessary standards are not in place yet in many cases and many of those that have been established are little used in the real world of large-scale, operationally deployed systems and products. Getting the appropriate standards defined, disseminated, and implemented in the marketplace is essential to progress in infrastructure.

There are five major groups of standards-developing organizations functioning today in areas relevant to information technology, electronic information and computer networking:

- International standards bodies, such as the International Organization for Standardization (ISO).

- National standards bodies in the US which link to the formal international organizations, such as the American National Standards Institute and its accredited standards writing bodies (for example, the National Information Standards Organization, NISO, which serves the library, publishing and information services communities).

- The National Institute of Standards and Technology (NIST; formerly the National Bureau of Standards, NBS) which develops standards for the federal government and also is charged to provide leadership in developing standards for the US generally in some situations where progress in standards is critical to US national interests and the private sector is not making sufficient progress.

- A growing array of ac-hoc industry standards development groups, consisting primarily but not exclusively of corporations; these are typically focused on a single problem. Examples include the UNICODE consortium, the Open Software Foundation (OSF), the Object Management Group, and many others.

- The Internet Engineering Task Force (IETF), an informal standards-writing group that manages standards for the Internet and is increasingly also concerned with developing standards to enable and facilitate the use of electronic information resources in the Internet environment.

There are major problems in the standards development system today. A full exploration of these is far outside the scope of this paper; the recent Office of Technology Assessment study Global *Standards: Building Blocks for the Future [U.S.* Congress, 1992], touches on a number of these problems but emphasizes the international perspective and standards in all areas, not just information technology and electronic information. Basically, from the perspective of building the networked information infrastructure, the speed with which formal standards (that is, standards within the ANSI/ISO structure and process) can be developed is too slow, leading to increased reliance on mechanisms like ad-hoc industry groups and the Internet Engineering Task Force. The costs for developing all types of standards have become very high; these high costs are largely precluding the effective participation of many of the communities involved in networked information in the standards development process. The refusal of the formal standards bodies to make their products available at reasonable cost and in electronic form has increasingly limited the usefulness of these products., particularly in disciplines like computer networking; by contrast, the IETF, which makes all of its work publicly available on the Internet, is gaining increased acceptance as a standards developer in many quarters, even though it is outside of the formal standards establishment. Finally, there is a growing perception among many of the people actually involved in building networks and the networked information infrastructure that the formal standards establishment has lost touch with engineering reality; the standards being developed by these groups are not being implemented in the marketplace and existing marketplace standards are not being reflected in the work of the formal standards bodies.[92] To some extent, at least, this problem is being created by conflicts between international demands, politics and commitments, and the policies of other nations regarding technology, standards, and the development of

92 This problem is perhaps most evident in the controversies surrounding the two competing networking standards suites: TCP/IP, which is the protocol that forms the basis of the Internet and is managed by the IETF, and is *not* a formal international standard, and the Open System Interconnection (0S1) protocol suite, which is a large and complex (and not yet complete) set of formal international networking standards that have been under development for about 15 years, but still have not gained large scale marketplace acceptance, despite attempts by various governments (including the US Government) to mandate their use. The history of this controversy is extremely complex and involves a number of political and economic as well as technical factors.

network infrastructure (and pressures of global marketplaces) and the requirements of the networked information community in the United States (which is certainly the dominant force in this area today) to build an effective infrastructure quickly and at reasonable cost. One consequence of this is that there is a considerable amount of at least partially duplicative work taking place, and the competing standards are causing confusion among the user community. There is also at least some anecdotal evidence that the private sector is increasingly turning away from the formal standards-setting process in frustration.

Some of the problems related to standards are funding problems; for example, it seems likely that relatively small investments, properly structured and applied, could help a great deal with the problem of broad community participation in networked information standards development and the speed with which such standards are developed. Encouraging marketplace implementation of standards is much more difficult; there is a fine art in developing and selecting standards that are viable and appropriate. To some extent this is the responsibility of purchasers, but today purchasers are often defeated in their attempts to acquire standards-conformant products by the lack of reasonable standards available to specify. Addressing the remain issues presents very complex policy and management problems for standards organizations and the communities that they serve, and, ultimately, for the United States government itself. A review of policies in the standards area as they relate to networking and networked information, with some specific emphasis on the relationship between policy choices and timely and effective progress on the construction of network and networked information infrastructure in the United States, is needed.

16. Recommendations for Possible Action

Legislative, Government and Public Policy Actions

- It is clear that cryptographic technology will be required to ensure the integrity of electronic information. Incorporation of this technology on a broad basis in the networked environment (including internationally) has been effectively paralyzed by a series of issues concerning intellectual property, standards, and exprot controls. It is time to address these issues in a systematic way and develop policies that will guide and encourage the implementation of the appropriate and needed integrity and authentication functions.

- Intellectual property issues are central to ensuring access and integrity for electronic information, particularly **as** this information becomes an increasingly substantial and important part of our intellectual record as a society. Consideration needs to be given both to clarifying the existing intellectual property laws as they relate to various forms of electronic information, particularly in a networked environment, and also to a review of whether the current balance between rightsholders and the public as defined in the copyright laws will continue to accomplish public policy objectives in the networked information environment. Immediate legislative action may well not be the answer; rather, the formation of a group similar to the National Commission on New Technological Uses of Copyrighted Works (CONTU) in the late 1970s might be an effective way to make progress at this time.

- A policy for networked access to federal government information which encourages such access at very reasonable costs is important both in its own right and also to stimulate technology development, network use, and to serve as a model for information policy development by governments at the state and local levels.

- A public policy debate related to expectations of privacy, confidentiality and anonymity in access to electronic information is needed to establish consensus on the objectives to be achieved by policy development and legislation. In addition, actions are needed to help make the public aware of the current state of affairs so that informed choices can be made by individuals. It may be appropriate to cast this effort in a larger context of individual privacy.

- Funding is needed to support research, development, prototype experiments, standards work, and networked information infrastructure construction. This is not the same as funding the development of the networks themselves. A substantial part of this need may be addressed by legislation currently under consideration by Congress. I believe that it is important to recognize that a substantial research component is needed here, and that a number of diverse, moderate scale prototypes will serve us better at this time than simply subsidizing the construction of one or two very large operational prototypes of "digital libraries".

- A major policy review of the standards development process and the organizations involved in this process is needed, in the context of information technology, electronic information, and computer-communications networks with consideration of both international implications and national needs. This should be conducted jointly with the standards development community and also seek broad participation by users of standards.

- A number of steps could be taken to ensure the public's access to electronic information resources. This might include some specific funding to help the existing library system, license of certain collections of information for unlimited use nationally, support to help schools license access to information, or other measures.

- Further consideration should be given to ways in which government funded electronic information resources, particular in the networked information environment, can help to achieve public policy goals such as controlling the increase in health care costs and improving the effectiveness of the educational system and national competitiveness. Such resources have proven effective in the past.

- Efforts to ensure the preservation of the cultural, historical and scholarly record as this becomes increasingly composed of electronic information are needed. This involves not just copyrighted information from publishers and other information providers but also public information. There are **a** number of government and nonprofit organizations with interests in this area, including (to name only a few) the National Endowment for the Humanities, the National Archives, the Library of Congress (both in its role as a library and as the registry for copyright), the Commission on Preservation and Access, the National Science Foundation, the National Library of Medicine, the National Library of Agriculture and the Association of Research Libraries. Consideration needs to be given to how to most effectively coordinate efforts on these issues.

Actions by the Stakeholders: Authors, Publishers, Libraries, Information Technology Providers and the Education Community

- There is a need for continued discussion among the stake holders as to the meaning of publication in the networked information environment and community expectations about continuity of access to electronic works, integrity of such works, and related topics. Organizations like the Coalition for Networked Information can play a key role in facilitating such discussions.

- A greater understanding of the increasing diversity of publication or information distribution paths available to authors and the implications of choices among these paths for the library, research and education, and publisher communities is needed. This may include some reassessment of the valuation placed on electronic publication channels in areas such as university tenure and promotion decisions, for example.

- Increased investment in the development and implementation of appropriate standards to facilitate the authoring, distribution and preservation of networked information is required. Further, these standards should reflect the evolving consensus about integrity and access to networked information.

- Publishers and libraries need to attempt again to reach a compromise about the uses of new technologies and their relation to copyrighted works which addresses the concerns of both parties. This could take the form of agreements about some limited' use of licensed electronic information for interlibrary loan, for example.

- Libraries, publishers and the scholarly community need to begin discussions about their roles and responsibilities in preserving and ensuring long-term access to the scholarly and cultural record. Broadcast media providers also need to be brought into these discussions. This is an area where it seems the financial stakes may not be high (unlike, perhaps, the interlibrary sharing of licensed electronic information) but which has great societal importance, and where it may be possible to make significant progress quickly. Such discussions might also provide a basis for informing future legislative options if appropriate.

Glossary

archie. Archie was developed at McGill University by Peter Deutsch and Alan Emtage, who have since established a private corporation (Bunyip) to further develop and commercialize the technology. In essence, archie automatically creates, manages, and offers access to a database of the contents of the major FTP archives worldwide and thus permits users to locate FTP archives holding files of interest to them. Note that archie is with a small “a”.

Ariel. Ariel is a system developed by RLG (which see) that permits documents to be faxed using the Internet rather than the dial telephone network as a transmission medium. Essentially, documents are scanned into a file on the sending host, transmitted across the network using FTP (which see) and then printed (or viewed on screen) at the receiving ARIEL node. ARIEL is based on IBM personal computer hardware platforms.

ARL. The Association of Research Libraries is a not-for-profit organization representing 119 major research libraries in the United States and Canada. ARL supports a wide range of activities, including studies of the function and costs of the research library system, the development and articulation of policy positions on legislation and other government activities that are of interest to the research library community, and development of resource sharing agreements and policies among its members. ARL is also one of the three parent organizations that created the Coalition for Networked Information (CNI).

BRS. BRS, which is now a part of the Maxwell communications empire, is another commercial search service very similar to Dialog (which see), although somewhat less extensive.

CARL. The Colorado Alliance of Research Libraries (CARL) is an organization based in Denver Colorado that started out building an online catalog for a number of Colorado based libraries, but has recently expanded nationally, offering a linked collection of catalogs for major libraries from Hawaii to Maryland. In addition, CARL creates and offers access to a database of journal tables of contents, a companion document delivery service, and a number of commercial databases.

Center *For Research Libraries.* The Center for Research Libraries (CRL) is an organization set up by a group of research libraries that essentially serves as a central depository for very rarely used material that is considered to be important to retain within the research library community but which is not used enough to justify any single library in this community retaining as part of the local collection; in these cases the material is sent to CRL where it can be available to the entire research library community but can be stored in a relatively inexpensive facility.

C/V/. The Coalition for Networked Information (CNI) is a joint project of EDUCOM and CAUSE, two associations concerned with information technology in higher education, and the Association for Research Libraries, an association of the 119 largest research

libraries in North America. Its purpose is to advance scholarship and intellectual productivity through the use of networks and information technology, and it has played an active role in promoting the development of networked information resources and the underlying technologies necessary to implement and use them.

CONTU. CONTU was the National Commission on New Technological Uses of Copyrighted Works which was established by Congress to make recommendations on copyright legislation in light of developments in technology. CONTU issued its final report in 1978. The work carried out by CONTU helped to define community standards for the interactions between technologies such as photocopying (xerography) and the copyright law, and to help define appropriate practices and standards for the acceptable use of these technologies in contexts such as Interlibrary Loan. CONTU also studied computer software and database intellectual property issues. CONTU made a number of recommendations for legislative changes; some were implemented and others were not. While the CONTU guidelines do not have the force of law, the represent a very real community consensus on acceptable behavior. In some cases they have been used by courts to help to interpret the copyright law, although other courts have ignored the CONTU recommendations as having no legal status.

Dialog. Dialog is a commercial online service that dates back to the late 1960s; originally a subsidiary of Lockheed, it is now owned by Knight-Ridder. Dialog essentially acts as a service bureau to database producers, mounting their databases and providing interactive searching access to them. It also handles billing and training, and typically pays royalties back to database producers who make their databases available through Dialog. Dialog is well known both for its very high prices, which have kept its use within the academic community to a minimum, and for its very complex user interface, which, while offering very powerful search capabilities, is really intended for use by trained searchers and not by end users. Dialog offers access to hundreds of databases; some of these are exclusively accessible though Dialog; others are also available though other channels (for example, CD-ROM, tape licenses direct to institutions, or other competing online services).

Fair Use. Fair use is a provision of the US copyright law which permits copying of copyrighted material for specific purposes, such as personal research. More generally, the term “fair use” is used to refer to the entire set of specific copying exemptions in the law, which include some provision for making a copy of an out of print work that is deteriorating for preservation purposes (if no other reasonable alternative is available), satire and criticism, and other exemptions.

FTP. FTP is the file transfer protocol within the TCP/IP protocol suite and widely used on the Internet to copy files from one host to another. It is the access mechanism used with FTP archives, which are simply large collections of files that are stored on various hosts on the Internet and available for copying. FTP supports an anonymous access mode that allows users to list and copy public files from many hosts on the network without any need for pre-registry with the archive host manager.

Gopher. Gopher is a system that was developed at the University of Minnesota in the early 1990s. It is a distributed system consisting of client programs for a variety of hardware platforms (including PCs, UNIX, and Macintoshes) and servers (again for

various platforms) which allow a server manager to establish a database of menus. The menu entries can point to other menus, either on the local Gopher server or any other Gopher server on the Internet, to documents (stored anywhere on the net), to interactive services, or various other types of information objects or resources. Essentially, Gopher offers a fairly simple, low cost means for an institution or even an individual to provide menu-based access to a variety of networked information (including locally stored information).

IETF. The Internet Engineering Task Force is a self-selected volunteer group that meets quarterly to develop standards for the Internet. While it is not an officially sanctioned national or international standards development body, it operates primarily by consensus and has been tremendously effective in managing the standards needed for the Internet environment. Standards developed by the IETF are called Requests for Comments (RFCs) for historical reasons. The IETF has a number of active working groups dealing with various aspects of standards and architectural models for networked information resources.

LISTSERV. LISTSERV (an abbreviation for list server) is a program that was originally developed for the IBM CMS operating system; more recently imitation programs with very similar functionality have been developed for the UNIX operating system environment. Essentially, the LISTSERV program permits mailing lists to be established with a wide range of parameters. Typically, users can join or leave a mailing list by sending commands to the LISTSERV program by electronic mail. Once they have joined a given mailing list, they can send a mail message to the list which will be echoed to all other subscribers of the list. The LISTSERV program also supports a variety of maintenance functions, such as maintaining and offering access to archived messages, permitting people to get lists of the people subscribed to a given mailing list, and the like. LISTSERV lists can also be defined as moderated, which means that the list moderator must approve all postings to the list before they are sent out to other subscribers.

Mail Reflectors. A mail reflector is essentially a simple, manual form of a LISTSERV (which see). Mail reflectors simply re-transmit mail to a list of users who are tabled as part of that mail reflector; in this sense a mail reflector can be viewed as simply a shorthand for sending mail to a list of people. Maintenance of the list is typically manual (perhaps assisted by computer programs) as distinct from the completely automated list management of LISTSERVs. In addition, there is usually no easy way to find out who is signed up on a given mail reflector.

Multicast. Multicast technology is a network facility which permits information to be broadcast to a specific subscriber group of machines attached to the Internet. Rather than having the source send one copy of the information to each recipient in the multicast group separately, multicasting permits the source machine to simply transmit one copy of the data, addressed to the multicast group address, onto the network; the network routing services take care of duplicating it as required so that every machine in the multicast group receives its own copy. Network hosts can join and leave a specific multicast group at will. Currently, multicast technology is used primarily on an experimental basis in the Internet (although it is a basic network service within some types of local area networking technology, such as Ethernet) to carry audio and/or

video traffic to groups of interested recipients in real time. Only a portion of the Internet (called the MBONE, or multicast backbone) supports multicast service at this time.

NNTP. The Network News Transfer Protocol is the protocol that is used by news readers that wish to access Usenet news groups from a newsgroup server. It is defined by RFC 997 [Kantor 1986].

OCLC. The Online Computer Library Center (formerly the Ohio Computer Library Center) is a not for profit organization based in Dublin, Ohio which provides a wide variety of services to the library community. These include interlibrary loan requesting and tracking, shared cataloging using an enormous copy cataloging database, and recently access to a variety of online databases (in a sense competing with commercial firms like Dialog (which see)). Recently OCLC has also partnered with the American Association for the Advancement of Science (AAAS) to mount the electronic journal Current *Clinical Trials,* and is currently working on several other electronic journals in conjunction with various other professional societies. The model for *CCT* is that of a database, where OCLC makes available the necessary user interface software to read articles, view them, and, optionally, print them, rather than the more common model used by other electronic journals where the contents of the journal are often physically transmitted to each subscriber.

PostScript. PostScript is a language developed by Adobe Systems to communicate with printers, particularly high-quality laser printers. Typically, word processors “print” a document by converting it into PostScript form; this is then sent to a printer for interpretation which produces the actual printed page. Programs also exist to preview PostScript files on display devices. It is very difficult to go backwards from PostScript to a revisable form document, or even one that permits reasonable full text searching; in addition, all semantic level markup (and even most syntactic level markup) is lost when a document is converted to PostScript form. In this sense, a PostScript document representation can be viewed as quite similar to a bitmapped image, although it is more compact.

Project *Gutenbeg.* Project Gutenberg, managed by Michael Hart, creates (either by scanning or keyboarding) ASCII versions of out-of-copyright books and other works, and makes these materials freely available over the Internet via anonymous FTP.

RLG. RLG is the Research Libraries Group, a consortium of research libraries in north America. One of the major activities of RLG is the operation of RLIN (the Research Libraries Information Network), a service similar to OCLC (which see) that provides libraries with access to a large shared cataloging database, an interlibrary loan requesting and tracking system, and a number of scholarly databases targeted for end users in academic institutions.

SGML. Standard Generalized Markup Language is an international standard for defining markup tagging for text. This can be at a relatively superficial level, where only syntactic structures such as headings and paragraphs are marked, thus facilitating the reformatting of a document for presentation in multiple environments (for example, in print or on a display terminal), or the markup can define very deep semantic meaning,

as is being doing in the Text Encoding Initiative project, which addresses the needs of the humanities computing community to encode text for subsequent computer analysis.

TELNET. TELNET is a protocol within the TCP/IP protocol suite and widely used on the Internet to conduct a terminal session with an interactive service on a remote host. It can be viewed as the network equivalent of dialing up a remote host with a character based terminal.

TOPNODE. This is a project of the Coalition for Networked Information (CNI - which see) to experiment with the description and cataloging of networked information resources.

Usenet *Newsgroups.* These are collections of electronic mail messages that are distributed throughout the Internet and beyond using the Usenet news system. There are many hundreds of such newsgroups, with more being established all the time. Total Usenet traffic is now measured in tens of millions of characters daily. Usenet news groups include the infamous “ALT” newsgroups, which are unmoderated and which deal with rather controversial issues such as sex and drugs, and which have been subject to censorship from time to time by various institutions. One subscribes to and reads Usenet news groups using a news reader. For a more extensive description of Usenet, see [Quarterman].

WA/S. WAIS stands for Wide Area Information Server; this is a system that is based on the **Z39.50** information retrieval protocol (which see) that was originally developed in the early 1990s by Thinking Machines, Apple, Dow Jones and KPMG. The original code was publicly distributed. Since that time a startup company (WAIS Inc.) has been established by several of the original developers to commercialize the technology, while work on the upgrading and extension of the public domain version continues through organizations like the Clearinghouse for Networked Information Discovery and Retrieval in North Carolina. Essentially, WAIS permits full text searching (using sophisticated statistical ranking and matching algorithms) against databases distributed across the Internet using a common user interface.

WWW. WWW is the WorldWide Web system originally developed by Tim Berners-Lee at the CERN center in Geneva, Switzerland. It can be viewed as a network-based hypertext system that allows linkages between arbitrary information objects stored on hosts throughout the network.

Z39.50. This is a US national standard developed by NISO, the National Information Standards Organization, an ANSI accredited standards developing body serving the publishing, library, and information services communities. The standard addresses computer to computer information retrieval and provides a basis for the development of network-based applications that offer a common interface to multiple, autonomously managed information servers.

Suggested Background Readings

For those unfamiliar with networking technology, networked information, and related issues the following brief list of suggested readings is provided. This list is not intended to be a comprehensive bibliography, even of survey works. Also included is a list of related Office of Technology Assessment reports that will provide helpful background on copyright issues, standards, cryptography, and other subjects related to the topic of this paper.

Networking Technology, Networks, and the Internet

Cerf, V. (1991). Networks. *Scientific American,* ***265(3),*** September 1991,42-51.

LaQuey, Tracy, with Ryer, J. C. (1993). The Internet *companion :* a *beginner's guide to global networking.* Reading, MA: Addison-Wesley.

Malamud, C. (1992). Stacks: *Interoperability in Today's Computer Networks.* Englewood Cliffs, NJ: Prentice-Hall.

Quarterman, J. S. (1990). *The matrix: computer networks and conferencing systems worldwide.* Bedford, MA: Digital Press.

Networked Information Services and Networked Information

American Society for Information Science (1992). *Networking, Telecommunications and the Networked Information Revolution, May 28-30, 1992, Proceedings of the ASIS 1992 Mid-Year Meeting. N.* Gusack (Ed.), Silver Springs, MD: American Society for Information Science.

Krol, E. (1993). *The whole Internet : user's guide & catalog* (Corrected Edition). Sebastopol, CA: O'Reilly & Associates.

Lynch, C. A., & Preston, C. M. (1990). Internet Access To Information Resources. in *Annual Review Of information Science And Technology, Volume* ***25 (1990),*** M. Williams (Ed.), pp 263-312.

Surveys of Topics Related to this Paper

U.S. Congress Office of Technology Assessment. (1987). *Defending Secrets, Sharing Data: New Locks and Keys for Electronic Information:* Washington DC: U.S. Government Printing Office.

U.S. Congress Office of Technology Assessment, (1986). *Intellectual Property Rights* ***in*** *an Age of Electronics and /formation.* Washington, DC: U.S. Government Printing Office.

U.S. Congress Office of Technology Assessment. (1988). *Informing the Nation: Fedeal /formation Dissemination in an Electronic Age.* Washington, DC: U.S. Governement Printing Office.

U.S. Congress Office of Technology Assessment. (1990). *Ctitical Connections: Communication future.* Washington, DC: U.S. Government Printing Office.

U.S. Congress Office of Technology Assessment. (1990). *Helping America Compete: The Role of Federal Scientific and Technical Information.* Washington, DC: U.S. Government Printing Office.

U.S. Congress Office of Technology Assessment. (1992). *Finding a* Balance: *Computer Software, Intellectual Property, and the Challenge of Technological Change.* Washington, DC: U.S. Government Printing Office.

U.S. Congress Office of Technology Assessment. (1992). Global *Standards: Building Blocks for the Future.* Washington, DC: U.S. Government Printing Office.

References

Alberti, B., Anklesaria, F., Linder, P., MaCahill, M., & Torrey, D. (1992). The Internet *Gopher protocol:* a *distributed document search and retrieval protocol.* Univeristy of Minnesota.

Arms, W. Y. (1992). The Design of the Mercury Electronic Library. *EDUCOM Review,* 27(6), 38-41.

Bagdikian, B. H. (1992). *The Media Monopoly* (4th cd.). Boston: Beacon Press.

Baker, S. K., & Jackson, M. E. (1992). *Maximizing Access, Minimizing Cost: A First Step Toward the Information Access Future.* Association of Research Libraries, Comittee on Access to Information Resources.

Balenson, D. (1993). *RFC 1423: Privacy Enhancement for Internet Electronic Mail: Part III: Algorithms, Modes and Identified*

Belkin, N. J., & Croft, W. B. (1992). Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Communications of the ACM, 35(12), 29-38.*

Berners-Lee, T. J., Cailliau, R., Groff, J.-F., & Pollermann, B. (1992). World-Wide Web: The Information Universe. *Electronic Networking: Research, Application and Policy, 2(l), 75-7.*

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM,* vol.28,(no.3), 289-99.

Bowers, F. J., & Shapiro, N. R. ***(1992).*** CD-ROM Standards: Essential for Progress. *CD-ROM Librarian, 7(8), 33-6.*

Brand, S. (1987). *The Media Lab. New* York: Viking.

Buckland, M. (1988). Bibliography, Library Records and the Redefinition of the Library Catalog. *Library Resources and Technical Services, 33(4), 299-311.*

Bull, G., Hill, l., Guyre, K., & Sigmon, T. (1991). Building an Academic Village: Virginia's Public Education Network. *Educational technology, 31(4), 30 (7* pages).

Burke, S. *(1991).* Los Angeles DA investigating Prodigy service. *PC Week, 8(18), 119 (2* pages).

Council on Library Resources (1990). *Communications in Supped of Science and Engineering: A Report to the National Science Foundation from the Council on Library Resources.* Council on Library Resources.

Council on Library Resources (1992). *Final Report on the Conference for Explorition of a National Engineering Information Service, June 14-19, 1992, Palm Coast, Florida.* Cornell Information Technologies and Media Services Printing.

de Sola Pool, 1. (1983). *Technologies of Freedom.* Cambridge, MA: The Belknap Press of Harvard University Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., & Landauer, T. K. (1990). Indexing by Latent semantic Analysis. *Journal of the American Society for /formation Science, 41(6), 391-407.*

Dillon, M. (1993). *Assessing Information on the Internet: Towards Providing Library Services for Computer Mediated Communication.* OCLC,lnc.

Emtage, A., & Deutsch, P. (1991). archie-an electronic directory service for the Internet. *Proceedings of the Winter 1992 USENIX Conference, 93-110.*

Foltz, P. W. (1990). Using Latent Sematic Indexing for Information Filtering. S/GO/S *Bulletin, 11(2-3), 40-7.*

Garfield, E. (1979). *Citation Indexing - Its theory and application in Science, Technology and Humanities. New* York: Wiley.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using Collaborative Filtering to Weave an Information Tapestry. *Communications of the ACM, 35(12), 61-70.*

Grunder, T. (1992). Whose Internet is it Anyway? - A Challenge. *Online Magazine,* 16(4), 6 (3 pages).

Grycz, C. (cd) (1992). *Serials Review Special Issue on New Mode/s in Serials Pricing; 18(1-2)*

Harman, D. (1992). The DARPA TIPSTER Project (Information Retrieval). *S/G/R Forum, 26(2), 26-8.*

Harman, D. (Ed.). (1993a). *The First Text Retrieval Conference (TREC-1)* (NIST Special Publication 500-207 cd.). Gaithersburg, MD: National Institute of Standards and Technology.

Harman, D. (1993 b). Overview of the First TREC Conference. In R. Korfhage, E. Rasmussen, & P. Willett (Ed.), *Sixteenth Annual International ACM S/G/R Conference on Research and Development in Information Retrieval,* (pp. 36- 48). Pittsburgh, PA, June 27- July 1 1993: ACM Press.

Jacso, P. (1992a). Author Agrees On Inspec Currency. *Database-The* Magazine *Of* Database *Reference And Review, 15(6), 55-55.*

Jacso, P. (1992 b). What Is In A(N) (Up)Date - Currency Test Searching Of Databases. *Database-The Magazine Of Database Reference And Review, 15(3), 28-33.*

Jacso, P. (1993a). A Proposal For Database Nutrition And Ingredient Labeling. *Database-The Magazine Of Database Reference And Review, 16(1), 7-9.*

Jacso, P. (1993 b). Searching For Skeletons In The Database Cupboard .1. Errors Of Omission. *Database-The Magazine Of Database Reference And Review, 16(1), 38-49.*

Jacso, P. (1993 c). Searching For Skeletons In The Database Cupboard .2. Errors Of Commission. *Database-The Magazine Of Database Reference And Review, 16(2), 30+.*

Kahle, B., Morris, H., Davis, F., & Tiene, K. (1992a). Wide Area Information Servers: An Executive Information System for Unstructured Files. *Electronic Networking: Research, App/lcations and Policy, 2(l), 59-68.*

Kahle, B., Morris, H., Goldman, J., Erickson, T., & Curran, J. (1992 b). Interfaces to Wide Area Information Servers. In N. Gusack (Ed.), *Networking, Telecommunications, and the Networked Information Revolution: ASIS 1992 Mid-year meeting,* Albuquerque, New Mexico: American Society for Information Science.

Kaliski, B. S. (1992). *RFC 1319: MD2 Message Digest Algorithm.*

Kaliski, B. S. (1993). RFC 1424: Privacy *Enhancement for Internet Electronic Mail:* Part *IV: Key Cetification and Related Services.*

Katz, A. R., & Cohen, D. (1992). *RFC* 1374: *File format for the Exchange of Images in the Internet.*

Kent, S. T. (1993). *RFC 1422: Privacy Enhancement for Internet Electronic Mail: Part II: Cetificate-Based Key Management.*

Keyhani, A. (1993). The Online Journal of Current Clinical Trials: An Innovation in Electronic Journal Publishing. *Database, 16(1), 14-15, 17-20, 22-3.*

Lederberg, J., & Uncapher, K. (1989). Towards a *National Colaboratoty: Report of an Invitational Workshop at the Rockefeller University, March 13-151989.* The Rockefeller University.

Lesk, M. (1991). The CORE electronic chemistry library. *S/G/R Forum,* 93-112.

Library of Congress (1991a). *Discussion Paper 49: Dictionary of Data Elements of Online Iformation Resources.* Washington, DC; Library of Congress.

Library of Congress (1991 b). *Discussion Paper 54: Providing Access to Online Information Resources.* Washington, DC; Library of Congress.

Library of Congress (1993). *Discussion Paper 69: Accommodating Online Systems and Services in USMARC.* Washington, DC; Library of Congress.

Library of Congress Network Advisory Committee (1992). Proceedings of the *Joint Meeting of the Library of Congress Network Advisory Committee and the Chief Officers of State Library Agencies Network* Planning Paper 23). Washington, DC; Library Of Congress.

Linn, J. (1993). RFC 1421: Privacy *Enhancement for Internet Electronic Mail: Part 1: Message Encryption and Authentication Procedures.*

Loken, S. C. (1990). *Report of the APS Task Force on Electronic /formation Systems.* Lawrence Berkeley Laboratory.

Love, J. P. (1993). The Ownership and Control of the US Securities and Exchange Commission's EDGAR System. *Government Publications Review, 20(1), 61-71.*

Lucier, R. E. (1990). Knowledge management: refining roles in scientific communication. *EDUCOM Review, 25,(3), 21-7.*

Lucier, R. E. (1992). Towards a knowledge management environment: a strategic framework. *EDUCOM Review, 27,(6), 24-31.*

Lynch, C. A. (1989). Library Automation and the National Research Network. *EDUCOM Review, 24,(3), 21-6.*

Lynch, C. A. (1991a). Visions of Electronic Libraries. In The *Bowker Annual Library and Book Trade Almanac, 36th Edition 1991* (pp. 75-82). New Providence, NJ: R. R. Bowker.

Lynch, C. A. (1991 b). The Z39.50 Information Retrieval Protocol: an Overview and Status Report. *Computer Communication Review, 21,(1), 58-70.*

Lynch, C. A. (1992). The Next Generation Of Public Access Information Retrieval Systems For Research Libraries - Lessons From 10 Years Of The Melvyl System. *Information Technology And Libraries, 11(4), 405-415.*

Lynch, *C.* A., Hinnebusch, M., Peters, P. E., & McCallum, S. (1990). Information Retrieval as a Network Application. *Library Hi Tech, 8,(4), 57-72.*

Lynch, C. A., & Preston, C. M. (1990). Internet Access To Information Resources. *Annual Review Of Information Science And Technology, 25, 263-312.*

Lynch, C. A., & Preston, C. M. (1992). Describing and Classifying Networked Information Resources. *Electronic Networking: Research, Applications and Policy, 2,(l), 13-23.*

Markoff, J. (1993, v 142 (Thursday, March 4, 1993)). Turning the desktop PC into a talk radio medium. (Carl Malamud's Internet Talk Radio Program). New York *Times.* p. Al(N), Al(L), COI 1.

Marshak, D. (1990). Filters: separating the wheat from the chaff. Patty Seybold's Office *Computing Report,* 13(11), 1-16.

Mayer, M. (1990). Scanning the Future (marketing use of supermarket scanning information). Forbes, 146(8), 114 (4 pages).

Mitchell, M., & Saunders, L. M. (1991). The virtual library: an agenda for the 1990s. *Computers in Libraries, 11(4), 8-11.*

Mitchell, W. J. (1992). *The Reconfigured Eye: Visual Truth in the Post Photographic Em.* Cambridge, MA: MIT Press.

Nelson, T. H. (1988). Managing Immense Storage: Project Xanadu provides a model for the possible future of mass storage. Byte, 13(1), 225-238.

Olson, M. V. (1993). The Human Genome Project. *Proceedings of the National Academy of Sciences of the United States of America, 90(10), 4388-4344.*

Palca, J. (1991). New Journal will Publish without Paper ('The Online Journal of Current Clinical Trials' Starts as an On-line Peer-reviewed Journal). *Science,* 253(5027), 1480.

Pertiz, B. C. (1992). On the objectives of Citation Analysis - problems of Theory and Method. *Journal of the American Society for nfonmation Science, 46(6), 448-451.*

Peters, P. E. (1992a). The Coalition for Networked Information as a Model. In J. W. Y. Smith (Ed.), ˆ(Eds.), *Networking and the Future of Libraries: Proceedings of the UK Office for Networding Conference,* (pp. 165-170). Bath, UK: Meckler.

Peters, P. E. (1992 b). Making the Market for Networked Information: An Introduction to a Proposed Program for Licensing Electronic Uses. *Setials Review,* 78(1-2), 19-24.

Piatetsky-Shapiro, *G., &* Frawley, W. J. (Ed.). (1991). *Know/edge Discovery in Databases.* Cambridge, MA: MIT Press.

Rivest, R. L. (1992a). *RFC 1227: MD5 Message Digest Algorithm.*

Rivest, R. L. (1992 b). *RFC 1320: MD4 Message Digest Algorithm.*

Roche, M. M. (1993). *ARL/RLG lnterlibrary Loan Cost Study˙ A Joint Effort by the Association of Research Libraries and the Research Libraries Group.* Association of Research Libraries.

Salton, G. (1988). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of /formation by Computer.* Reading, MA: Addison-Wesley.

Saylor, J. M. (1992). NEEDS (The National Engineering Education Delivery System): If We Build It (According to Standards) They Will Come! In *Networks, Telecommunications and the Networked Information Resources Revloution. American*

Society for Information Science, *Mid-year Proceedings.,* (pp. 273+). Albuquerque, NM: American Society for Information Science.

Schwartz, M. F. *(1989).* The networked resource discovery project. */formation Processing 89. Proceedings of the IFIP Ilth World Computer Congress, 827-32.*

Schwartz, M. F., Emtage, A., Kahle, B., & Neuman, B. C. (1992). A Comparison of Internet Resource Discovery Approaches. *Computing Systems, 5(4), 461-93.*

Schwartz, M. F., Hardy, D. R., Heinzman, W. K., & Hirschowitz, G. C. (1991). Supporting resource discovery among public Internet archives using a spectrum of information quality. *11th International Conference on Distributed Computing Systems* (Cat. *N0.91CH2996-7), 82-9.*

Simmonds, C. (1993). Searching Internet Archive Sites with archie: Why, What, Where and How. *Online, 17(2), 50 (5* pages).

Stonebraker, M. (1992). An Overview of the Sequoia 2000 Project. In *Digest of Papers. COMPCON Spring 1992. Thirty-Seventh IEEE Computer Society International Conference,* (pp. 383-8). San Francisco, CA: IEEE Computer Society Press.

Stonebraker, M., & Kemmitz, G. (1991). The POSTGRES Next - Generation Database Management System. *Communications of the ACM, 34(10), 78-92.*

Stout, C. (1992). *TENET: Texas Education Network* Texas Education Agency, Austin.

Strangelove, M. (1993). *Directory of Electronic Journals, Newsletters, and Scholarly Discussion Lists.* Association of Research Libraries.

Tenopir, C., & Ro, J. S. (1990). *Full Text Databases.* Westport, CT: Greenwood Press.

U S. Congress Office of Technology Assessment. (1987). *Defending Secrets, Sharing Data: New Locks and Keys for Electronic Information.*

U.S. Congress Office of Technology Assessment. (1986). *Intellectual Property Rights in an Age of Electronics and Information.* Washington, DC: U.S. Government Printing Office.

U.S. Congress Office of Technology Assessment. (1990). *He/ping America Compete: The Role of Federal Scientific and Technical Information.* Washington, DC: U.S. Government Printing Office.

U.S. Congress Office of Technology Assessment. (1992). Global *Standards: Building Blocks for the Future.* Washington, DC: U.S. Government Printing Office.

Vinge, V. (1992). *A Fire upon the Deep. New* York: TOR.

Watkins, B. (1991). USC to Put Full Text of The Chronicle on network. *Chronicle of Higher Education, 37(27), A20.*

Watkins, B. T. (1992). Free-Net helps Case Western fulfill its community-service mission. Chronic/e *of Higher Education,* 38(34), A21 (2 pages).

White, H. S. (1989). The vaule-added process of librarianship. *Library Journal, 114(1), 62 (2* pages).

Wiggens,R. (1993). Gopher. *Public Access Systems Review. 4(l).*