NORTH ATLANTIC TREATY ORGANISATION

RESEARCH AND TECHNOLOGY ORGANISATION

AC/323(IMC-002)TP/06

www.rta.nato.int

RTO LECTURE SERIES PRE-PRINTS **IMC-002 (2004)**

Electronic Information Management

(La gestion électronique de l'information)

Edited by Yaşar Tonta

The material in this publication was assembled to support a Lecture Series under the sponsorship of the Information Management Committee (IMC) presented on 8 - 10 September 2004 in Sofia, Bulgaria.

Published September 2004

Distribution and Availability on Back Cover

NORTH ATLANTIC TREATY
ORGANISATION

RESEARCH AND TECHNOLOGY
ORGANISATION

AC/323(IMC-002)TP/06

www.rta.nato.int

RTO LECTURE SERIES PRE-PRINTS **IMC-002 (2004)**

Electronic Information Management

(La gestion électronique de l'information)

Edited by Yaşar Tonta

The material in this publication was assembled to support a Lecture Series under the sponsorship of the Information Management Committee (IMC) presented on 8 - 10 September 2004 in Sofia, Bulgaria.

The Research and Technology Organisation (RTO) of NATO

RTO is the single focus in NATO for Defence Research and Technology activities. Its mission is to conduct and promote co-operative research and information exchange. The objective is to support the development and effective use of national defence research and technology and to meet the military needs of the Alliance, to maintain a technological lead, and to provide advice to NATO and national decision makers. The RTO performs its mission with the support of an extensive network of national experts. It also ensures effective co-ordination with other NATO bodies involved in R&T activities.

RTO reports both to the Military Committee of NATO and to the Conference of National Armament Directors. It comprises a Research and Technology Board (RTB) as the highest level of national representation and the Research and Technology Agency (RTA), a dedicated staff with its headquarters in Neuilly, near Paris, France. In order to facilitate contacts with the military users and other NATO activities, a small part of the RTA staff is located in NATO Headquarters in Brussels. The Brussels staff also co-ordinates RTO's co-operation with nations in Middle and Eastern Europe, to which RTO attaches particular importance especially as working together in the field of research is one of the more promising areas of co-operation.

The total spectrum of R&T activities is covered by the following 7 bodies:

- AVT Applied Vehicle Technology Panel
- HFM Human Factors and Medicine Panel
- IST Information Systems Technology Panel
- NMSG NATO Modelling and Simulation Group
- SAS Studies, Analysis and Simulation Panel
- SCI Systems Concepts and Integration Panel
- SET Sensors and Electronics Technology Panel

These bodies are made up of national representatives as well as generally recognised 'world class' scientists. They also provide a communication link to military users and other NATO bodies. RTO's scientific and technological work is carried out by Technical Teams, created for specific activities and with a specific duration. Such Technical Teams can organise workshops, symposia, field trials, lecture series and training courses. An important function of these Technical Teams is to ensure the continuity of the expert networks.

RTO builds upon earlier co-operation in defence research and technology as set-up under the Advisory Group for Aerospace Research and Development (AGARD) and the Defence Research Group (DRG). AGARD and the DRG share common roots in that they were both established at the initiative of Dr Theodore von Kármán, a leading aerospace scientist, who early on recognised the importance of scientific support for the Allied Armed Forces. RTO is capitalising on these common roots in order to provide the Alliance and the NATO nations with a strong scientific and technological basis that will guarantee a solid base for the future.

Published September 2004

Electronic Information Management

(RTO-LS-IMC-002 (2004) Pre-Prints)

Executive Summary

Networked information sources and services proved to be an indispensable part of our everyday lives. We get access to a wide variety of bibliographic, full-text and multimedia databases through the intranets, extranets and the Internet. Information and communication technologies available in different settings (e.g., workplace, home, library, and Internet cafés) facilitate our access to such online services as e-banking, e-government, e-learning, and e-entertainment. Although some advanced information processing and networking capabilities are available to us, we still experience difficulties in searching, finding, gathering, organizing, retrieving, and using information. Trying to find information among billions of electronic sources is likened to trying to "drink water from a fire hydrant." We need well-designed electronic information management systems and services to better manage information that we use in our private and professional lives. Availability of such systems and services are of paramount importance to all organizations large and small.

Electronic information management can be defined as the management of information that is recorded on printed or electronic media using electronic hardware, software and networks. It includes the description of strategies, processes, infrastructure, information technology and access management requirements as well as that of making economic, legal and administrative policies with regards to the management of electronic information.

This publication aims to review current developments in electronic information management. It contains ten papers covering a wide range of topics. The titles of papers are as follows: "Internet and Electronic Information Management," "New Initiatives for Electronic Scholarly Publishing: Academic Information Sources on The Internet," "Information Discovery and Retrieval Tools," "Electronic Collection Management and Electronic Information Services," "Economics of Electronic Information Provision," "Metadata for Electronic Information Resources," "Preservation of and Permanent Access to Electronic Information Resources," "Electronic Information Management and Intellectual Property Rights," "Infrastructure of Electronic Information Management" and "The Digital Library – The Bulgarian Case."

Papers explore several trends, models, and strategic, operational and policy issues with regards to electronic information management. Among them are: customization and personalization of electronic information services; Davenport's ecological model of information management; current developments in electronic journals, electronic prints, electronic theses and dissertations; initiatives to create a global network of archives of digital research materials (e.g., Budapest Open Access Initiative); features and capabilities of search engines; use of metatags to describe contents of electronic documents; electronic collection management strategies and models (e.g., "access versus ownership" and "pay-per-view"); economics of preparing and providing published information; alternative models of electronic information provision; metadata and resource discovery; digital preservation and archiving projects; intellectual property rights in the digital information environment; the European Union directive on copyright and information society; access control devices (e.g., finger-printing and time-stamping); and networking infrastructure for electronic information management; and the digital library initiatives in Bulgaria.

The material in this publication was assembled to support a Lecture Series under the sponsorship of the Information Management Committee (IMC) and the Consultant and Exchange Programme of RTA presented on 8 - 10 September 2004 in Sofia, Bulgaria.

Table of Contents

[†] Not available at the time of printing

List of Authors/Lecturers

Prof. Dr. Yaşar TONTA
Lecture Series Director
Department of Information Management
Hacettepe University
06532 Beytepe
Ankara
TURKEY

Ms. Bonnie CARROLL
Information International Associates, Inc.
1009 Commerce Park Dr., Ste 150 /
P.O. Box 4219, Oakridge, TN 37830
UNITED STATES

Mr. Graham CORNISH
Copyright Circle
33 Mayfield Grove, Harrogate
North Yorkshire HG1 5HD
UNITED KINGDOM

Prof. Dr. Ana Maria R. CORREIA
UNL/ISEGI
Campus Campolide
1070-124 Lisbon
PORTUGAL

Ms. Gladys A. COTTER
USGS/BRD
Mail Stop 302
12201 Sunrise Valley Drive
Reston, VA 22092
UNITED STATES

Mr. Michael T. FRAME
USGS Center for Biological Informatics
Mail Stop 302
12201 Sunrise Valley Drive
Reston, VA 22092
UNITED STATES

Ms. Gail HODGE
Information International Associates, Inc.
312 Walnut Place
Havertown, PA 19083
UNITED STATES

Ms. Andrea JAPZON
Information International Associates, Inc.
1009 Commerce Park Dr.
Suite 150 / P.O. Box 4219
Oakridge, TN 37830
UNITED STATES

Dr. Dincho KRASTEV
Central Library
Bulgarian Academy of Sciences
1, "15 Noemvri" St.
1040 Sofia
BULGARIA

Prof. Dr. José Carlos TEIXEIRA
Departamento de Matematica
Universidade de Coimbra
Largo D. Dinis - Apartado 3008
3001-454 Coimbra
PORTUGAL

Mr. Gregory D. TWITCHELL
USGS/BRD Customer Support Center
Mail Stop 302
12201 Sunset Valley Drive
Reston, VA 22092
UNITED STATES

Acknowledgements

The IMC Committee wishes to express its thanks to the organisers from Bulgaria, for the invitation to hold this meeting in Sofia, and for the facilities and personnel which make the meeting possible.

Introduction

Yaşar Tonta
Editor and Lecture Series Director
Hacettepe University
Department of Information Management
06532 Beytepe, Ankara
TURKEY

tonta@hacettepe.edu.tr

Bibliographic, full-text and multimedia databases available through the intranets, extranets and the Internet are of paramount importance to all organizations large and small. Networked information services proved to be an indispensable part of every day lives of users working for both commercial and non-profit organizations as well as of more casual users with personal interests to pursue. Almost half a billion people try to get access to networked information sources and services every day. More often than not they are confronted with too much information. Although search engines, "knowbots," and "intelligent agents" are of some use in this area, trying to find information among billions of electronic sources is likened to trying to "drink water from a fire hydrant." Well-designed electronic information management systems and services can facilitate users' task and enable them to better cope with too much information in their private and professional lives.

Organized by the Information Management Committee (IMC) of the Research and Technology Organization (RTO) of NATO for the Partnership for Peace (PfP) Nations, Lecture Series on Electronic Information Management aims to review current developments on electronic information management. It explores a wide variety of operational and policy issues with regards to electronic information management ranging from available sources and services to the description, organization, management, preservation and archiving of electronic information collections, to infrastructure, economics and intellectual property rights of electronic information provision.

Available to participants prior to the Lecture Series, this book provides background information on the topic and can serve as an additional resource to support the lectures. It contains 10 papers of lecturers on various aspects of electronic information management. References to both printed and electronic information sources listed in each paper can be useful. Full-text of papers including links to cited sources will be made available through the Web site of IMC (http://www.rta.nato.int). What follows is a brief overview of each paper in the order of their appearance in the book.

In the first paper, "Internet and Electronic Information Management," Dr. **Yaşar Tonta** reviews the latest developments in the electronic information scene. He draws attention to the amount of information produced annually in the world (about five exabyte), increasing processing, storage and transmission capacities of computers as well as declining costs of computer hard drives and network bandwidths. He discusses issues of electronic information description and organization in detail along with development and management of electronic information collections. He reviews the developments in information technologies that gave way to customization and personalization of electronic information services. Dr. Tonta emphasizes the importance of preserving and archiving electronic information and speculates whether publishers, information centers and aggregators would assume this responsibility. Dr. Tonta then goes on to introduce Davenport's ecological model for information management. In his ecological model Davenport sees information, its collection, description, organization, management, and use in a broader context and thinks that information can be better managed if we take its three interrelated and interdependent environments into account, namely (1) the more immediate "information

environment," which consists of the whole set of cross-relationships among information people, strategies and policies, processes, technology, information culture and behavior; (2) the "organizational environment," and (3) the "external environment." Dr. Tonta points out that Davenport's ecological model is also applicable to electronic information management.

In their paper, "New Initiatives for Electronic Scholarly Publishing: Academic Information Sources on The Internet," Professors **Ana Maria Ramalho Correia** and **José Carlos Teixeira** review the evolution of scientific communication and discuss in detail the current developments in electronic journals, electronic prints, electronic theses and dissertations (ETDs) and other digital collections of "grey literature" (i.e., technical reports). They provide several examples of such repositories containing electronic information sources that can be used for academic research. The Los Alamos Physics Archive, providing access to e-prints of some 200.000 articles on high-energy physics; and the Networked Digital Library of Theses and Dissertations (NDLTD), providing free access to graduate theses and dissertations in a distributed environment, are among them. Correia and Teixeira also describe initiatives to create a global network of archives of digital research materials (e.g., Open Archives Initiative, Budapest Open Access Initiative) and discuss such policy issues as prior publication, and the possible roles of library and information professionals in self-publishing schemes and in creating digital archives of ETDs.

In his paper, "Information Discovery and Retrieval Tools," **Michael T. Frame** reviews the principles of how search engines work by means of a model. He describes the ways in which search engines discover the existence of Web documents and provides a list of metatags that are used most frequently by search engines for discovery and indexing. He also touches upon the issue of "spam," which some Web site developers are inclined to use to falsify search engines so that their content will be indexed more favorably by search engines and retrieved before the other sites in the retrieval output. His paper ends with a list of features and capabilities of search engines along with some recommendations to content and software developers to improve the discovery and retrieval of their content.

In their paper, **Gladys Cotter**, **Bonnie Carroll**, **Gail Hodge** and **Andrea Japzon** provide a comprehensive overview of electronic collection management and electronic information services. After a brief discussion on the digital revolution that is currently taking place in library and information centers, they first tackle the issue of electronic collection management and review the major collection management strategies. They identify the key challenge in collection management as being that of "ownership vs. access" and stress that the move to electronic information management is transforming information centers to "access-based organizations." They discuss the issues of selection, acquisition, cataloging, and archiving of electronic information in detail. Next, authors review the electronic information services and concentrate on electronic reference, information delivery, and education of users and personnel. They conclude that electronic collection management and electronic information services are in a period of rapid transition, and the technology used to manage the information allows for extensive innovation in information selection, description, distribution, retrieval, and use.

In his paper, "Economics of Electronic Information Provision," **Graham Cornish** covers the economics of preparing and providing published information. He examines the role of different "players" in the publishing chain including authors, editors, publishers, distributors, and users. He also does this for the provision of electronic information and reviews the roles of libraries. He challenges the view that libraries are supermarkets and argues that libraries are not solely run on the basis of commercial motives and that their purchasing policy is not dictated by commercial needs. Libraries make strenuous efforts to collect materials for all types of users and they do not discourage certain types of users such as the children and the elderly. He considers the question of who will pay for those unable to afford access. Finally, he discusses alternative models of electronic information provision (e.g., SPARC, the Scholarly Publishing and Academic Resources Coalition) and reviews the roles of licensing consortia such as ICOLC (International Coalition of Library Consortia) and JSTOR (Journal Storage Online).

Gail Hodge reviews the issues with regards to metadata for electronic information resources. She points out that the rationale for creating metadata remains the same for both electronic and printed resources (to facilitate resource discovery and access), although the terminology has changed (from cataloging and indexing to "metadata"). She describes the purpose of metadata (to discover, locate and organize electronic information resources) and the methods by which metadata can be created (manual vs. through metadata editors and generators). She provides a basic metadata structure and summarizes the characteristics of major metadata schemes (Dublin Core, GILS, TEI, and EAD, to name a few). She also discusses the issues of "metadata interoperability" among different schemes and the importance of controlled vocabularies for subject indexing of electronic information sources.

In her second contribution, entitled "Preservation of and Permanent Access to Electronic Information Resources," **Gail Hodge** starts with the definitions of basic terms such as "digital archiving," "digital preservation" and "long-term access" and gives an outline of major projects including JSTOR, InterPARES (International Research on Permanent Authentic Records in Electronic Systems) and ERPANET (Electronic Resources Preservation and Access Network). She offers a framework for archiving and preservation of electronic information and addresses a number of issues comprehensively. Among them are the creation and acquisition of electronic information, collection development, metadata and archival storage formats for preservation, migration and emulation, access, rights management and security requirements. She also discusses the emerging stakeholder roles and identifies key issues in archiving and preservation of electronic information such as long-term preservation and intellectual property rights.

In his second contribution, **Graham Cornish** addresses the intellectual property rights in the context of electronic information management. He clarifies the use of such basic terms as "copyright," "copy," "author," "publisher," "user" and "fair use" in the digital environment. He explains access control devices including fingerprinting, watermarking, and stamping, and gives examples of their use in the European Union (EU) projects such as CITED (Copyright in Transmitted Electronic Documents) and COPY SMART. He discusses the impact of the latest EU directive on copyright and information society and the complexities of implementing this directive in different legal regimes and cultural environments.

Gregory D. Twitchell and **Michael T. Frame**'s paper addresses the infrastructure of electronic information management. Using a non-technical language as much as possible, they describe the following tools and technologies: network infrastructure, mass storage devices, JAVA, proxy servers, network address translation, firewalls, tunnelling, forwarding, encryption, and routing. They emphasize that the key to a robust, flexible, secure, and usable network systems is to establish a strong network infrastructure and point out that network hardware and applications are co-dependent. They conclude that a standard component of a good network management is the planning and review process, and organizations must take a proactive role in this process to make sure that they have a secure, reliable, usable and scalable network.

The book concludes with Dr. **Dincho Krastev**'s piece on the development of digital libraries in Bulgaria. Dr. Krastev provides a summary of some of the early digital library projects and gives a detailed description of the digitization of Slavic manuscripts that was carried out in the Central Library of the Bulgarian Academy of Sciences in cooperation with the national and international institutions. He ends his paper by emhasizing the importance of team work to succeed in such collaborative digitization projects involving specialists in Slavic manuscripts and medieval texts, computational medieval studies and computational humanities.

Internet and Electronic Information Management

Yaşar Tonta
Hacettepe University
Department of Information Management
06532 Beytepe
Ankara
TURKEY

tonta@hacettepe.edu.tr

ABSTRACT

The number and types of information sources accessible through the Internet are ever increasing. Billions of documents including text, pictures, sound, and video are readily available for both scholarly and every-day uses. Even libraries and information centers with sizable budgets are having difficulties in coping with this increase. More developed tools and methods are needed to find, filter, organize and summarize electronic information sources. This paper is an overview of a wide variety of electronic information management issues ranging from infrastructure to the integration of information technology and content, from personalization of information services to "disintermediation." It discusses the issues of description, organization, collection management, preservation and archiving of electronic information and outlines Davenport's "ecological model" for information management and its components, namely, strategy, politics, behavior and culture, staff, processes, and architecture.

1.0 INTRODUCTION

Lyman and Varian (2003) estimates the amount of new information produced in the world in 2002 to be around five exabytes (one exabyte = one billion gigabytes, or 10^{18} bytes). "Five exabytes of information is equivalent in size to the information contained in 37,000 new libraries the size of the Library of Congress book collections" and "the amount of new information stored on paper, film, magnetic, and optical media has about doubled in the last three years" (Lyman and Varian 2003). Printed documents of all kinds constitute only .01% of the total whereas information recorded on magnetic media such as personal computers' hard disks constitutes an overwhelming majority (92%) of the overall information production. Lyman and Varian (2003) point out that "almost 800 MB of recorded information is produced per person each year" and they "estimate that new stored information grew about 30% a year between 1999 and 2002" (see Table 1).

Table 1: Worldwide Production of Original Information, if stored digitally, in Terabytes circa 2002. Upper estimates assume information is digitally scanned, lower estimates assume digital content has been compressed.

Storage Medium	2002 Terabytes Upper Estimate	2002 Terabytes Lower Estimate	1999-2000 Upper Estimate	1999-2000 Lower Estimate	% Change Upper Estimates
Paper	1,634	327	1,200	240	36%
Film	420,254	76,69	431,690	58,209	-3%
Magnetic	5187130	3,416,230	2,779,760	2,073,760	87%
Optical	103	51	81	29	28%
TOTAL:	**5,609,121**	**3,416,281**	**3,212,731**	**2,132,238**	**74.5%**

Source: Lyman and Varian (2003).
Available: http://www.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm#summary

We are faced with an enormous rate of increase of electronic information. For comparison, library collections double every 14 years whereas the *annual* growth rate for information available through the Internet was about 300% during the early years, although there are some signs that the growth rate of the public web is leveling off in the last couple of years (O'Neill, Lavoie and Bennett 2003). The Library of Congress, one of the largest libraries in the world, has accumulated some 170 million items over decades. Yet the number of documents on the "surface Web" was estimated to be about 2.3 billion in 2000 (Bergman 2000). The number of documents on the "surface web" increased tremendously since then. In June 2004, one search engine (Google) alone performs its searches on more than 4.2 billion documents (http://www.google.com). The total number of documents should be close to 10 billion documents. Bergman pointed out that if one included dynamically created web pages, documents and databases available through the enterprise intranets, the number of documents went up to 550 billion! Although accessible through the Web, such documents and databases are usually behind the firewalls and therefore not directly accessible through the regular search engines (hence called "deep Web") (Bergman 2000). In size, Lyman and Varian (2003) measured the volume of information on the surface web about 167 terabytes as of Summer 2003. They estimate the volume of information on the deep web as somewhere between 66,800 and 91,850 terabytes!

As more information sources are born digital (or later become digital) and publicly accessible through the Internet, the relative importance of the management of information in personal, organizational and societal levels also increases tremendously. This makes the management and retrieval of information from large quantities of electronic sources all the more important. We are expected to know how to discover, find, filter, gather, organize, store, and get access to recorded information. We need to manage information successfully and be avid "consumers" of information so as to successfully manage our professional and personal lives. Alvin Toffler (1992) warns us that the ignorants of the future are not going to be the ones who do not know how to read and write ("illiterates") but those who do not know how to find and get access to relevant information (so called "information illiterates"). Considered by many people as the next stage of "literacy", "information literacy" includes abilities to handle hardware and software (computers, networks, Web, etc.) used to find information as well.

Several philosophers and scholars including, among others, Plato, Bacon, and Wells have given, considerable thought to knowledge, classification, and information retrieval problems. Some discussed issues with regards to the recording, storage and retrieval of information. Plato, for instance, raises what is

called "Meno's Paradox" and draws attention to the difficulties of searching for knowledge. The following dialogue between Meno and Socrates reflects this:

> "MENO. But how will you look for something when you don't in the least know what it is? . . even if you come right up against it, how will you know that what you have found is the thing you didn't know?
>
> "SOCRATES. . . . Do you realize that . . . a man cannot try to discover either what he knows or what he does not know? He would not seek what he knows, for since he knows it there is no need of the inquiry, nor what he does not know, for in that case he does not even know what he is to look for." (*Plato's Meno* 1971: 31-32).

An interesting argument so far as it goes. Although the argument is mainly philosophical and was discussed by Plato in the context of "virtue" vs. "knowledge," searching for knowledge should not be that different from searching for information, after all. If Socrates is right, "[o]ne can never find out anything new" (*Plato's Meno* 1971: 6), nor can search for or know about anything. So, we shall not explore it further. (For more information on Meno's Paradox, see Evans 1995.)

2.0 INFORMATION AND COMMUNICATION TECHNOLOGIES

In the past, some people were sceptical about the value of information and communication technologies (ICT) at the beginning. What follows are a few examples reflecting such scepticism:

- "Who needs this [telephone] invention? We have a lot of little boys to carry messages." Chief Engineer at the American Post Office, 1876.
- "Each town may wish to have one telephone." U.S.A PTT Director General, 1886.
- "Telephone is not something that would interest millions. It is a facility for rich people; it is a commercial tool for those who could afford it." *Times*, 1902.
- "I think that as many as five computers would be sold all over the world." Thomas Watson of IBM, 1943.
- "In the future computers would weigh as little as 1.5 tons." *Popular Mechanics*, 1949.

Despite early scepticism, ICT have always played a paramount role in information processing and management. Such technologies constitute the infrastructure that "makes it possible to store, search, retrieve, copy, filter, manipulate, view, transmit, and receive information" (Shapiro and Varian 1999: 8). ICT products enable us to perform all the abovementioned activities. Shapiro and Varian (1999: 84-85) draw attention to the fact that digital technology sharply reduces *both* copying and distribution costs of information by dramatically reducing the cost of making perfect reproductions and by allowing those reproductions to be distributed quickly, easily, and cheaply. No other technology has succeeded reducing *both* reproduction and distribution costs so far.

Postman (1993: 4-5) pointed out that ". . . it is a mistake to suppose that any technological innovation has a one-sided effect. Every technology is both a burden and a blessing; not either-or, but this-and-that." Technology gets cheaper and cheaper to produce and distribute information. Yet, it increases the volume of available information and creates what is called "information overload." Reuters, the British News Agency, is producing 27.000 pages of documents per second. Users get inundated with information that they do not need, are unable to digest or simply have no time to "process." Existence of too much information creates what is called "analysis paralysis" (Waddington 1997). As the late Nobel laureate Herbert Simon put it, "a wealth of information creates a poverty of attention" (Shapiro and Varian, 1999: 6). In other words, it is not enough to provide just the content: it is much more important and difficult to "attract the eyeballs" of the potential users (Reich 2002: 42). In addition to the technology for producing

and distributing information, we need technology to help simplify processing of information by the users. Otherwise, as Varian (1995: 161-162) points out, "[t]echnology for producing and distributing information is useless without some way to locate, filter, organize and summarize it."

Cost of storing and transmitting information is also sharply decreasing while the storage capacities and capabilities of various magneto-optic storage and transmission media are ever-increasing. For instance, the cost of storing data on computer hard disks went down a couple of order of magnitudes in a decade (Fig. 1). A similar trend can also be observed in the cost of transmitting data through the networks of various types (Fig. 2). Yet, such networks allow us to transmit large volumes of data in seconds. For example, a couple of years ago Nortel Networks Corporation developed a new product that "uses 160 channels, and each channel can transmit 10 billion bits of information a second. The new system could transmit 1.6 trillion bits a second over a single optical fiber." This may not mean much to those of us who are not familiar with network speeds. To put it in context, then, Nortel points out that this transmission capacity is "enough . . . to transmit the contents of the Library of Congress across the country in 14 seconds" (Schiesel 1999).

Figure 1: Hard Drive Cost per Gigabyte.

Source: Lyman and Varian (2000).
Available: http://www.sims.berkeley.edu/research/projects/how-much-info/charts/charts.html

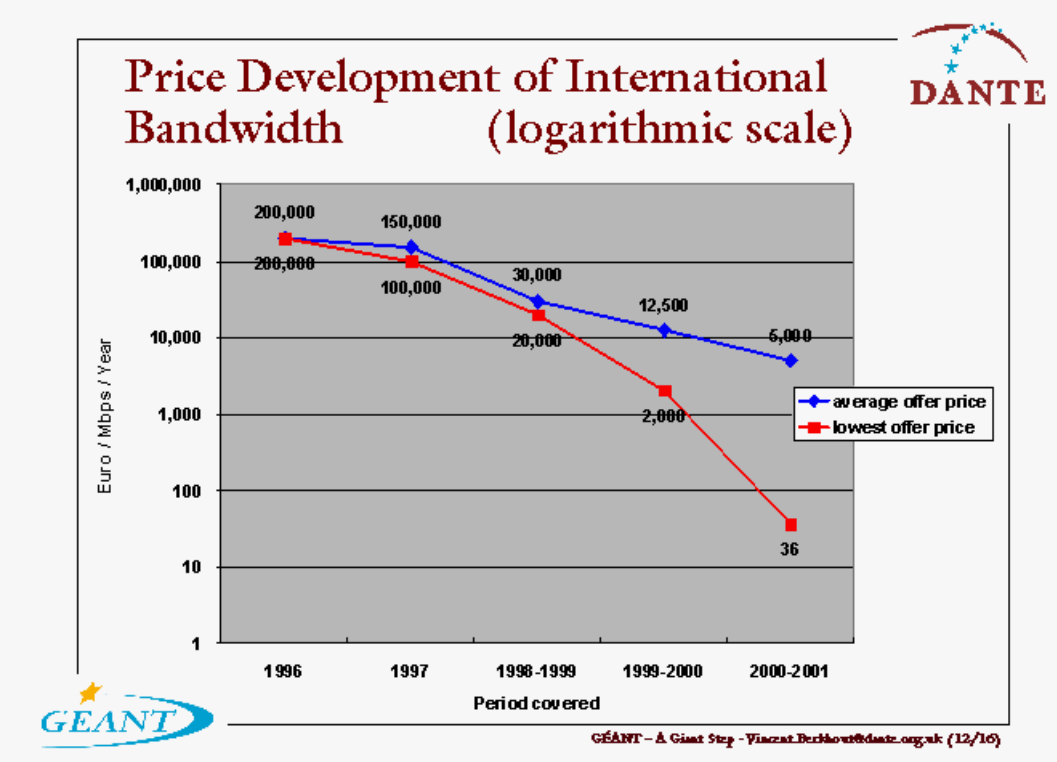


Figure 2: Price Development of International Bandwidth (Logarithmic Scale).

Source: Berkhout (2001).
Available: http://www.dante.net/geant/presentations/vb-geant-tnc-may01/sld012.htm

As Varian (1995) points out, information technologies developed in recent years allow us to store all the recorded information produced by humankind throughout the centuries on a computer chip and carry it in our pockets. Moreover, such a chip containing the cumulative depository of knowledge can be implanted in our heads and used as an extension of human brain. According to George B. Dyson, the "globalization of human knowledge," as H.G. Wells prophesied some 60 years ago, is about to become a reality as many distributed databases are made available through the Internet:

> The whole human memory can be, and probably in a short time will be, made accessible to every individual . . . . This new all-human cerebrum . . . need not be concentrated in any one single place, it need not be vulnerable as a human head or a human heart is vulnerable. It can be reproduced exactly and fully in Peru, China, Iceland, Central Africa, or wherever else seems to afford an insurance against danger and interruption (Dyson 1997: 10-11).

Thanks to the latest developments in information technologies, the accumulated knowledge of centuries can now be distributed quickly and easily.

3.0 INFORMATION DESCRIPTION AND ORGANIZATION

Storing and transmitting large amounts of information, replicating it in different places over the globe, or even making it an extension of human brain for collective intelligence still requires speedy access to and retrieval of useful information. For this, information needs to be organized. "Information to be organized needs to be described. Descriptions need to be made of it and its physical embodiments" (Svenonius, 2000: 53). The description of information is then the *sine qua non* of both organization and, consequently, retrieval of information. Svenonius maintains that:

> There is an essential difference between organizing information to compile an encyclopedic compendium of knowledge and organizing it for the purpose of information retrieval. In the former, what is ordered and arranged is the information itself; in the latter, it is the documents embodying information (such as books systematically arranged on library shelves) or their surrogates (such as catalog cards alphabetically arranged in a catalog). In the context of information retrieval, the modus operandi of information organization is not compilation but description (Svenonius 2000: 206).

It may at first seem an easy task to describe electronic documents (web sites, logs of discussion lists, etc.) embodying information so as to organize and later retrieve them. Yet this is one of the most difficult tasks in the Internet environment. The transient nature of Web documents sometimes makes it impossible to discover and describe information sources available through the Internet. Some Web documents are dynamically created "on-the-fly." Some ephemeral electronic documents such as meeting announcements simply disappear automatically once they fulfil their functions. In general, the average life of a Web document is estimated to be 44 days (Kahle 1997: 82-83). It is likely that some electronic documents are removed from the Web before they get noticed and described by the search engines and/or human indexers. Moreover, both the information itself (content) and its "metadata" get lost forever if it is not discovered and described before it disappears. In printed documents the information source and the metadata describing its contents are usually separated. If the printed document gets lost, one still has its metadata. This is unlike in electronic documents where both content and its metadata usually come together. Losing the document usually means losing its description as well. Although search engines are of some help in this area, they do not necessarily index all the documents available on the Web. In fact, any given search engine indexes only a fraction (e.g., 16% for Northern Light) of all the Web documents (Guernsey 1999).

We should stress the fact that describing documents is not a mechanical process. Search engines usually index documents on the basis of existence of certain keywords in the documents. They do not necessarily try to relate similar documents to one another. In fact, this is the main difference between machines and the human brain as to how they organize information. The human brain organizes information by means of what Vannevar Bush called "associative indexing" (Bush 1945: 101-108). Even though the two pieces of information are not described with the same subject keywords, the human brain can still make connections between them.

Search engines can "describe" electronic documents simply by using some statistical techniques (e.g., the frequency of keywords). Yet, there is more to the description of documents. First, search engines use different term weighting algorithms to extract keywords from documents. Consequently, they weight and rank the same documents differently. Similarly, human indexers tend to assign different index terms to the same document. In other words, the consistency of indexing has been quite low even among professional indexers (Tonta 1991).

Second, and perhaps more importantly, it is sometimes impossible to have an agreed-upon definition of certain terms for various reasons. For instance, despite the efforts of international organizations such as the United Nations since 1970s, the term "terrorism" cannot be defined to the satisfaction of all the Nations, which "has been a major obstacle to meaningful international countermeasures" (UNODC 2004).[1] More currently, Nations bordering with the oil-rich Caspian Sea cannot decide whether it is a "sea" or a "lake" in view of conflicting economic stakes. If one does not have agreed-upon definitions of certain terms, how can then one describe documents satisfactorily on terrorism or the Caspian "Sea"?

Third, classification of documents, which comes after description, is another difficult issue. In his book, *Women, fire and dangerous things,* George Lakoff discovered that Australian aborigines classify women,

[1] For the history of the dispute and the European Union's definition of terrorism, see Dumitriu (2004).

fire and dangerous things together for some reason (Lakoff 1990). One cannot decide easily if "alternative medicine" can be categorized under the subject of health sciences, religion, philosophy or all of the above (Rosenfeld and Morville 1998: 24). We cannot even make up our minds if tomato is a vegetable or a fruit.[2]

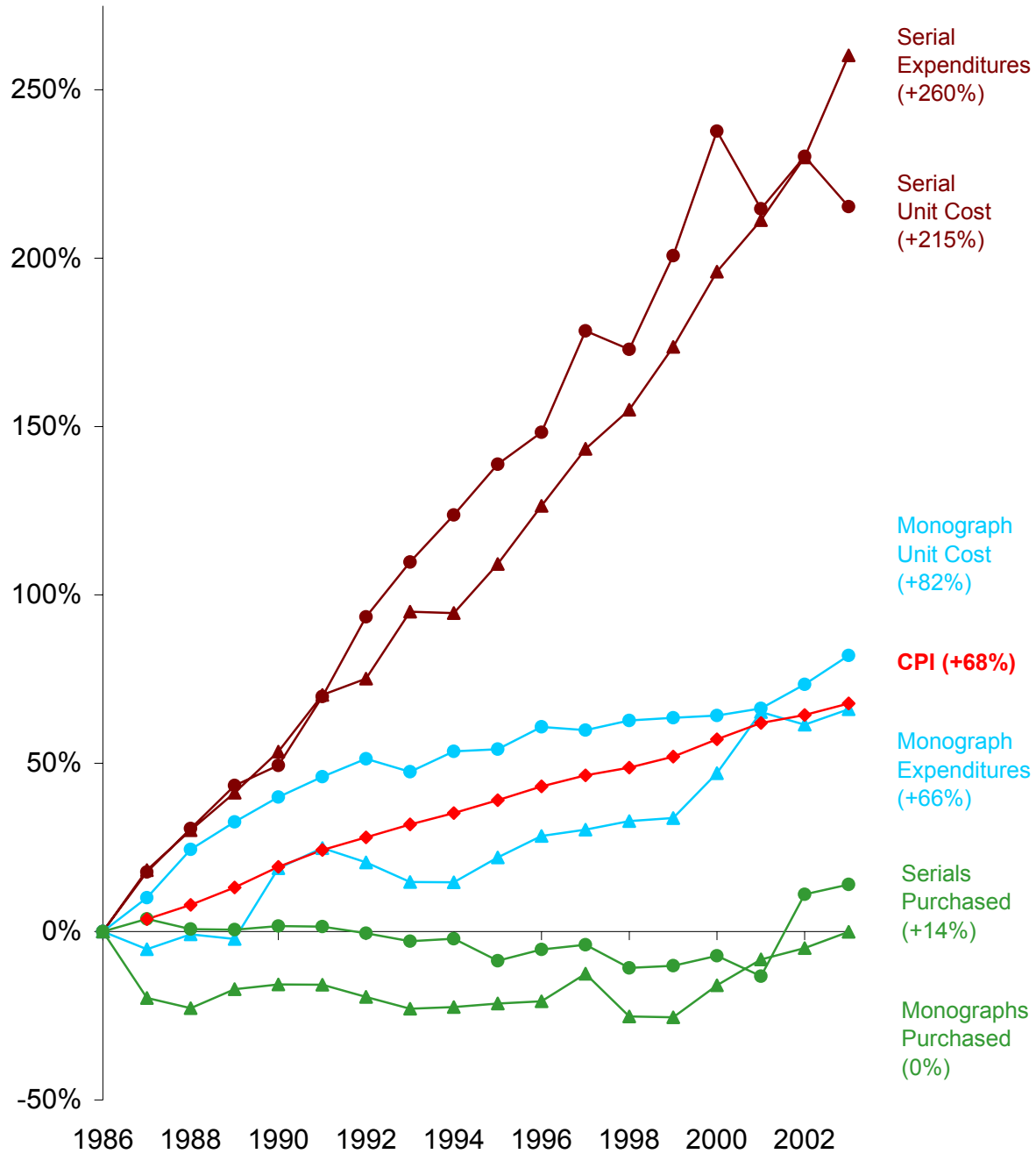
4.0 COLLECTION MANAGEMENT OF ELECTRONIC INFORMATION

In traditional libraries, owning a book or a journal guarantees access to the contents of that information source at least by one user. If the library does not own the source, then its users have to look for it elsewhere. This is not the case for networked information sources. To provide access to such sources through the Web is as valuable as owning them on site for information centers. In light of the availability of information sources from remote locations for simultaneous use, collection development and management has become a more crucial activity in information centers. Information centers are no longer limited with their own resources to provide information services to their users. Developments of Web-based information sources and services have partly coincided with exorbitant increases of prices of printed information sources, especially scientific journals, which further encouraged information managers to try new approaches. For instance, the number of serial titles subscribed to by members of Association of Research Libraries (ARL) has increased only 14% during the period of 1986-2003, whereas the expenditures for serials has increased 260% in the same period (Fig. 3).

[2] Regarding "tomato" being a vegetable or a fruit, Rosenfeld and Morville (1998: 24), cite an interesting case (from Grady 1997) that the U.S. Supreme Court had tried in the past: "The tomato is technically a berry and thus a fruit, despite an 1893 U.S. Supreme Court decision that declared it a vegetable. (John Nix, an importer of West Indies tomatoes, had brought suit to lift a 10 percent tariff, mandated by Congress, on imported vegetables. Nix argued that the tomato is a fruit. The Court held that since a tomato was consumed as a vegetable rather than as a desert like fruit, it was a vegetable.)"

Graph 2

Monograph and Serial Costs in ARL Libraries, 1986-2003



Source: *ARL Statistics 2002-03*, Association of Research Libraries, Washington, D.C.

Figure 3: Monograph and Serial Costs in ARL Libraries, 1986-2003.

Source: http://www.arl.org/stats/arlstat/graphs/2003/graph2_03.xls

The Internet made it possible to try new financial approaches to better manage electronic information "collections." "Ownership vs. access" is one such approach.[3] Owning information sources "just in case" users might need them is no longer the dominant method of collection development in libraries and information centers. Instead, information centers concentrate on providing "just in time" access to electronic sources should the users need them. Ownership vs. access approach enables libraries to get access to more resources. At the same time, expenditures for processing, maintenance and storage of information sources get reduced so that more money could be spent on license fees of electronic information sources.

Vendors move from subscription-based economic models to some uncertain ones to sell information or license its use. In addition to ownership vs. access, there are a number of other approaches that could be applied in electronic collection management. Pay-per-view, transaction-based pricing, per-access charges, individual and institutional licenses or combinations thereof are among them.

It should be noted that the availability of information sources through various economic models put additional burden on collection managers. It is no longer sufficient to buy or subscribe to information sources, process, maintain and store them. Collections of information centers are not limited with what they own, maintain and archive. Separate policies of processing, maintenance, storage and usage need to be developed for different sources licensed or acquired through certain channels. Information managers need to develop policies for sources that are usually maintained and archived by other agencies. They have to develop policies for sources that they get access to through mirror sites. For instance, the electronic library of the University of California at Berkeley (http://sunsite.berkeley.edu) classifies electronic information sources under four groups: (1) Archived: Sources that the library own and are committed to permanent archiving as well as providing continued access; (2) Served: Sources that are maintained by the library, yet no decision has been made for their permanent archiving; (3) Mirrored: Sources that are maintained by other agencies and mirrored by the library; and (4) Linked: Sources living on remote computers, yet the library provides links to them from its own Web site (Digital, n.d.).

Naturally, responsibilities of the information centers for electronic sources under each category differ considerably from each other. For instance, in addition to institutional commitment, powerful computers, large data warehouses and bandwidth may be needed to process, store and transmit information under the first two categories (Archived and Served).

Networked information services are gradually becoming the most heavily used services in most library and information centers. What should library and information centers do to cope with this increasing demand? What types of changes, if any, should be expected in the organizational and administrative structures of library and information centers due to networked services?

It appears that the trend towards providing just in time access to more and more electronic information sources alters the use patterns of information centers. Users can conveniently consult the bibliographic databases from their own desktops located at their labs, dorms, or homes. They can download the full-texts of articles from electronic journals which their library has a license for. They can request electronic copies of books and articles from collaborating libraries or they can place online document delivery requests

[3] In fact, the trend towards access rather than ownership has been observed in other walks of life, too. The relative importance of owning property has been decreasing. For instance, people prefer to buy "experience" of different types of vacation packages rather than buying summer houses. In his book, *The age of access: how the shift from ownership to access is transforming modern life*, Jeremy Rifkin stresses that in the new economy wealth and success are measured not in terms of ownership of physical capital (plants, materials, etc.) but in terms of control of ideas in the form of intellectual and intangible capital. Ideas and talent are more important than plant and material. Rifkin points out that: "The new information-based industries – finance, entertainment, communications, business services, and education – ready make up more than 25 percent of the U.S. economy. . . . The information sciences. . . are based less on ownership of physical property and more on access to valuable information, be it embedded in software or wetware" (Rifkin 2000: 52-53). Rifkin briefly touches upon the "ownership vs. access" debate taking place in the library world (p. 87-88).

(if other libraries do not have access to those resources, either). These and some other automated transactions can be completed without even paying a visit to the library. Just in time access increases user satisfaction and provides "instant gratification." This is reflected in the service trends in ARL libraries: the number of circulation and reference transactions has decreased for the first time in recent years while the interlibrary borrowing has almost doubled during the last decade (Kyrillidou and Young, 2001) (see Fig. 4). Consequently, library and information centers spend ever-increasing percentages of their total budgets to the provision of electronic information sources. For instance, the average percentage of the total materials budget devoted to electronic resources by U.S. research libraries was 13.5% in the academic year of 2000/2001. Some libraries (e.g., Johns Hopkins, Texas, and Arizona Universities) spent more than 20% of their total budget on electronic resources (Sewell 2001).

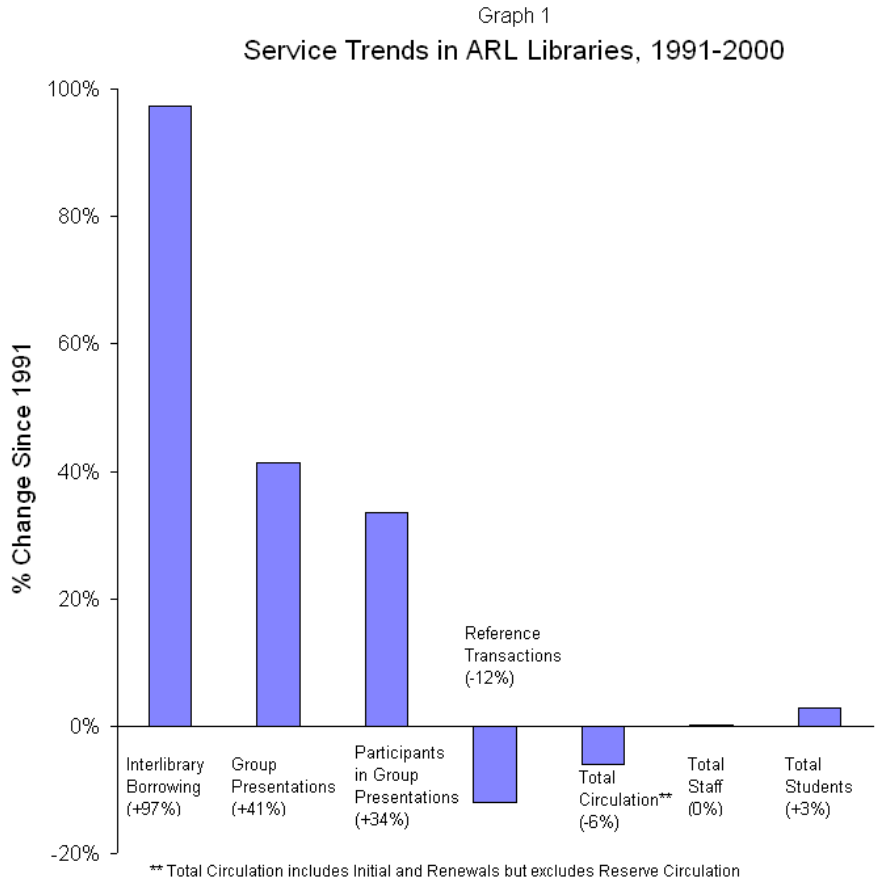


Figure 4: Service Trends in ARL Libraries 1991-2000.

Source: Kyrillidou and Young (2001, graph 1).
Available: http://www.arl.org/stats/arlstat/graphs/2000t1.html

Information managers are faced with the challenge of providing better services with shrinking budgets. Collection managers of information centers are getting together to provide consortial access to electronic sources to get more favourable deals from information vendors. Interlibrary cooperation and coordination of resource sharing is facilitated by the network environment as it is easier for information centers to form consortia and share electronic information sources. Although traditional resource sharing arrangements encouraged competition rather than cooperation in view of the benefits that large libraries accrued by owning research materials, this is no longer the case as small libraries and information centers can get access to information sources over the network with the same speed as the large ones can, regardless of where the sources are held. Furthermore, as indicated above, introduction of new pricing models by publishers such as licensing (rather than subscription) and access fees for electronic information sources

and relatively favourable offers for consortial agreements has made the economics of cooperation more visible. This does not necessarily mean that information managers have all the wherewithal to tackle such as access management, long-term preservation and archiving of electronic information in consortial collection development schemes. Nevertheless, in terms of satisfying information needs of their users, they are in a much better position in the new environment than they were before (Tonta 2001: 292-293).

5.0 MASS CUSTOMIZATION AND PERSONALIZATION OF ELECTRONIC INFORMATION GOODS AND SERVICES

The term "mass customization" in the title of this section may sound a bit odd as we are used to reading about "mass production" or "mass distribution" in the "mass media." As Robert R. Reich, former U.S. Secretary of Labor, points out in his book, *The future of success: working and living in the new economy*, mass production has been the pillar of the Industrial Era as it enables us to produce the same goods in large quantities in assembly lines, thereby reducing the cost of each item produced. He cites Henry Ford's famous quote "Any customer can have a car painted any color that he wants, as long as it's black" and emphasizes that "Henry Ford's assembly line lowered the cost and democratised the availability of automobile, by narrowing choice" (Reich 2002: 16).

Standardized goods and services of mass production have been highly regarded in industrial societies. The futurist Alvin Toffler, on the other hand, has first mentioned the importance of "unstandardized" goods and services for the "society of the future" in 1970s in his famous book, *Future shock* (Toffler, 1970: 234-235). Developments in computer-aided manufacturing, computer and network technologies have proved Toffler right. It is now possible to produce goods exactly as you want them (which is called customization or personalization) at the best price and highest quality. Moreover, goods can be ordered from anywhere in the world as distance is no longer a constraint. In view of these three characteristics of production ("as you want them", "from anywhere", "at the best price and highest quality"), Reich labels this era as "The Age of the Terrific Deal" (Reich 2002: 13-26).

Customized production of a wide variety of goods and services in an economy is seen as an indication of a rich and complex society. As Toffler (1970: 236) emphasized: ". . . pre-automation technology yields standardization, while advanced technology permits diversity." A wide variety of customized goods and services are available in several industries: computer, automotive, textile and manufacturing industries, hotels, airlines and health services, to name just a few.

Hart (1995) defined mass customization by using two distinct definitions:

1) The visionary definition: The ability to provide customers with anything they want profitably, any time they want it, anywhere they want it, any way they want it.
2) The practical definition: The use of flexible processes and organizational structures to produce varied and often individually customized products and services at the low cost of a standardized, mass production system (cited in Mok, Stutts and Wong, 2000).

Organizational structures of companies involved in mass production and mass distribution differ from those involved in mass customization. Mass production and mass distribution is based on mechanistical organizational structures where the main objective is to produce more of the same products in large quantities and cheaper than their competitors. Administrative structure is hierarchical. Continuous development is rewarded. The training is traditional. Mass customization, on the other hand, requires dynamic organizational structures where the main objective is to produce what the customer exactly wants. The idea is not to sell more of the same products to different customers once and make more profits but to attract more customers and keep them satisfied (as it is six times more expensive to find new customers than keeping the returning ones). Administrative structure is flattened. Customer-oriented continuous training is the norm.

The use of advanced information technologies makes it cheaper to produce personalized goods and services. Toffler points out that ". . . *as technology becomes more sophisticated, the costs of introducing variations declines*" (1970: 236; italics in original). This is also true of information goods and services: on-demand publishing of textbooks, readers, and newspapers, online book stores (e.g., Amazon.com), personalizing news portals (e.g., MyCNN, MyYahoo!), personalizing banking, health, education, and travel services are some of the diverse services that came into being due to advanced networking technologies.

Electronic information services are increasingly becoming personalized within the last decade. Software packages such as MyLibrary enable users to customize and personalize their electronic information environments. They define searches and identify sources that they use most frequently (including search engines, reference sources and electronic journals) and automatically get regular search results, current awareness and table of contents (TOCs) services. They get personalized document delivery and user education services. Companies try to make reference services available on every desktop computer using software (e.g., http://www.liveperson.com) that enables them to have reference queries answered live without even setting a foot in the library.

Electronic information services such as access to electronic journals, reference and document delivery services are gradually becoming integrated with automated library systems (Tonta 2003). For instance, to provide seamless access to electronic information sources and services, rights and privileges of each user can easily be defined depending on his/her status (i.e., student, faculty, remote user). Automated library systems can then recommend certain resources based on user's interests and privileges. Such "recommender systems" have been in use for some time by search engines and online book stores. Data gathered through the analysis of users' past interactions and transactions with the system (sites visited, books bought, etc.) are used to develop (mass) personalized information systems and recommend products and services that might conceivably be of interest to individual users. Electronic information centers also collect such valuable data about their users as well as use of their collections and services. It appears that library and information centers are reluctant to introduce recommender systems based on users and use data, presumably because they are trying to tackle issues of privacy, economics of information, and electronic payments of royalties, to name just a few.

Commercial companies producing personalized goods and services are going through a restructuring process so that they can adapt to change. They become less centralized with fewer hierarchical levels and respond to user needs more quickly. So are information centers providing electronic information services. They, too, gradually switch from centralized model of information management to the distributed one. Economic models based on centralized information management tend to produce information goods and services on the basis of "one size fits all" approach whereas models based on personalization approach users with the understanding that an ongoing relationship will be built. Personalization aims to recognize users when they use the services and provides personal help and advice if and when needed.

6.0 DISINTERMEDIATION

Personalization of electronic information services may sound like establishing a more "personal" relationship with the users. This is not the case, however. Information about each user (his/her characteristics, habits, use patterns, etc.) is gathered by electronic means. Users sometimes provide information about themselves voluntarily by filling out forms or accepting cookies. Sometimes they have to supply information in order to get privileges of access to services. Or, sometimes aggregates of users are identified from transaction logs of system use. As a result of disappearance of face-to-face communication with users, information professionals are no longer able to "intermediate" between the users and the resources (Tonta 2003).

It appears that one of the consequences of the applications of information technology is what is called "disintermediation." Reactions towards disintermediation are mixed. Some are against it solely because they lose personal touch while others see it as an opportunity to cut costs and "reintermediate" with remote users. Take banking, for example. A face-to-face transaction carried out within the building costs banks as much as six times more of what an online transaction does. Furthermore, customers usually have to wait in the line and pay for some of those face-to-face services (e.g., money orders) while they are instantly available and free of charge if carried out through the Internet. Who will miss, then, the personal touch of, and an opportunity to exchange pleasantries with, an already overloaded teller?

Similar arguments can be put forth for personalized electronic information services as well. Face-to-face transactions cost information centers more compared to their electronic equivalents that are available through the Web. For instance, a university library comparing the costs between Web-based resources and those on its internal CD-ROM network found that average cost per search made on Web-based databases was as low as 15 cents for some databases while the average cost per search on all CD-ROMs was 15 dollars (Lindley 2000: 334). Similar comparative figures are available for the average cost of downloading an article from electronic journals as opposed to providing it through document delivery services.

Brown and Duguid (2000: 21-22) criticize the over reliance on information and point out that it leads to what they call "6-D vision." The 6-D vision consists of what authors call six "futurist-favored words" starting with "de-" or "dis-" including "disintermediation" (other five being demassification, decentralization, denationalization, despacialization, and disaggregation). They think that:

> First, the evidence for disintermediation is far from clear. Organizations . . . are not necessarily becoming flatter. And, second, where it does occur, disintermediation doesn't necessarily do away with intermediaries. Often it merely puts intermediation into fewer hands with a larger grasp. The struggle to be one of those few explains several of the takeovers. . . . It also explains the "browser wars" between Netscape and Microsoft, the courtship of AT&T and Microsoft, and the continuing struggle for dominance between Internet Service Providers (ISPs). Each of these examples points not to the dwindling significance but to the continuing importance of mediation on the 'Net (as does the new term *infomediary* . . .). Moreover this kind of limited disintermediation often leads to a *centralization* of control (Brown and Duguid 2000: 28, italics original).

We also witnessed such mergers and takeovers in online information industries in the last couple of years. Yet, it remains to be seen what effect, if any, they will have on the restructuring of information centers providing online information services.

7.0 PRESERVATION AND ARCHIVING OF ELECTRONIC INFORMATION

Continued access to information is only possible through preservation and archiving. Preservation and archiving of electronic information sources differ significantly from that of printed sources. Preserving physical media on which information is recorded (such as books and journals) guarantees preserving the intellectual content in them. One can get access to information unless the physical medium is not damaged (Graham 1994). The first e-mail message did not survive; no "documentary" record exists for it today. Satellite observations of Brazil in the 1970s got lost as they were recorded on now obsolete tapes.

Preservation and archiving of information stored on electronic media is quite problematic. First, the life of electronic media is relatively short compared to the more traditional media such as paper and microfiche (magnetic media: 10-30 years, optical disks: 100 years, paper: 100 years, microfilm: 300 years). Second, information-bearing objects are nowadays usually "bundled" with the technology by which their content can be deciphered. In other words, one needs computer, communication and network technologies

that run multimedia software to read, hear, and view electronic documents. Thus one has to buy not only the content but also the technology as well. Furthermore, both content and the technology should be preserved so as to get access to information.

Preservation and archiving of electronic information is based on "copying." Information recorded on old media needs to be transferred to the new media from time to time so that it will not get inaccessible due to the obsolescence of technology. New technologies do not necessarily supplant the old ones. Information published on different media (paper, microfiche, CD-ROM, etc.) will co-exist for quite some time. The copying process for preservation and archiving purposes is called "technology refreshment" or "technology migration" (Preserving 1996). Information recorded on more traditional media can also be copied onto electronic media ("digitisation"). Needless to say, migrating information from one medium to another creates several formatting problems. New versions of a software package (used with the newer technology) may sometimes not recognize the information (e.g., footnotes) that was prepared using earlier versions of the same package. Manes (1998) offers some practical guidelines for preservation and archiving of electronic information: using simple formats for copying, preserving image files without compression, using the same software for both creation and archival of electronic documents, archiving two copies of each document using good quality media, developing an archival plan before upgrading hardware and software, and, testing if new hardware is able to read the archival copies. In other words, storage, back up, refreshment and access mechanisms needed to preserve and archive electronic information should be seen as long term investments. Organizations managing electronic information need sound IT support to integrate contents with computer and communication technologies.

While the responsibility of preserving and archiving printed information rests on libraries and archives, it is not clear who is responsible for the preservation and archival of electronic information. Just as publishers were never held responsible for preservation of printed documents in the past, it is likely that they will not be relied upon for the archival copies of electronic documents, either. Some publishers refrain from even saving electronic copies of their own titles. As for-profit ventures such as publishers or "aggregators," the decision to preserve and archive is mainly shaped by commercial motives.

It is not yet clear, though, if libraries will be solely responsible for the preservation and archival of electronic information sources for various reasons. First, as we pointed out earlier, it gets cheaper to store large volumes of electronic information. Electronic information sources occupy much less space. The key question here is, of course, to decide what to preserve and what to discard. As this has been a most difficult decision to make in libraries, archives and museums, some even contemplate of preserving everything just because it is cheaper doing so. It is likely that some institutions (publishers, commercial companies, non-profit organizations such as author guilds) may wish to assume the preservation and archival of electronic information in view of economic feasibility.

Second, libraries and archives preserving printed sources do not usually get due credit for this task. They do not necessarily reap the benefits of owning those resources as the use is limited to one person only at any given time and location. The situation is quite different for institutions archiving electronic sources: they can provide access to those resources through the Internet regardless of time and location. This makes it attractive for publishers to get involved in archival business. Whereas storage and archival of printed sources consumes additional expenses, immediate access to archival copies of electronic books and journals through the publishers' Web sites generates new source of income for publishers.

Third, some publishers and non-profit organizations assume the role of long term archiving of electronic information sources published by themselves as well as other publishers. JSTOR, for instance, aims to be a reliable and long term archive of scientific journals and makes the electronic copies of those journals available through the Web. Institutions such as national libraries and universities try to become repositories of electronic copies of dissertations and technical reports. Although it remains to be seen what role information centers are to play in the preservation and archiving of electronic information sources,

the responsibility will be assumed by more than one institution in a distributed environment, as Hedstrom (1998) indicated:

> . . . societies only allocate a small and finite amount of resources to preserving scholarly and cultural resources. And in the digital environment it seems likely that more preservation responsibilities will be distributed to individual creators, rights holders, distributors, small institutions, and other players in the production and dissemination process.

Preservation and archiving of electronic information that is based on "copying" also engenders heated discussions on issues of access and copyright for electronic information. The authenticity and integrity of copied information becomes more difficult to ascertain. More research is needed to uniquely identify digital objects for description as well as for electronic commerce and royalty payments by means of electronic copyright management systems (ECMSs). Such unresolved issues make digital preservation a "time bomb" for electronic information management (Hedstrom 1998).

8.0 AN ECOLOGICAL MODEL FOR ELECTRONIC INFORMATION MANAGEMENT

Electronic information management can be defined as the management of information that is recorded on printed or electronic media using electronic hardware, software and networks. It includes the description of strategies, processes, infrastructure, information technology and access management requirements as well as making economic, legal and administrative policies with regards to the management of electronic information.

In his book, *Information ecology: mastering the information and knowledge environment*, Thomas H. Davenport takes a more holistic approach to information management. His approach is based on ecological paradigm, which sees information in relation to its environment. What follows is a detailed review of his views on an ecological model of information management that he developed. Needless to say, his general approach to managing information is also applicable to managing electronic information.

Davenport (1997: 28) points out that most companies and organizations have done two things to manage information better: "They've applied technology to information problems, and attempted to use machine-engineering methods to turn data into something of use on computers." Davenport stresses the fact that a more holistic approach is needed and offers the "information ecology model." The following quotation summarizes his views better:

> Information ecology includes a much richer set of tools than that employed to date by information engineers and architects. Information ecologists can mobilize not only architectural designs and IT but also information strategy, politics, behavior, support staff, and work processes to produce better information environments. When managers manage ecologically, they consider many avenues for achieving information objectives. They rely on disciplines of biology, sociology, psychology, economics, political science, and business strategy – not just engineering and architecture – to frame their approach to information use. And they look beyond a company's immediate information environment to the overall organizational environment – how many buildings, offices and physical locations are involved? What kind of technology is already in place? What is the current business situation? – as well as the external market environment.
>
> . . . there are four key attributes of information ecology: (1) integration of diverse types of information; (2) recognition of evolutionary change; (3) emphasis on observation and description; and (4) focus on people and information behavior (Davenport 1997: 28-29).

Davenport gives the example of Amazon rain forest as a physical ecology and points out that in the overall ecology of a rain forest, there are three environments that are interconnected: treetops, shadowy world under the leaves, and soil underground. Change in one environment will result in changes in other environments. Davenport sees three environments in any information ecology as well: information environment, organizational environment, and external environment. He thinks that:

> To date, there are no practically oriented approaches that encompass all components of an information ecology – that is, how an aggregate of individuals, in a particular organization, in a particular industry affected by broader market trends, works with, thinks about, focuses on, and generally manages information. . . But description is a fundamental attribute of information ecology. And to manage ecologically, we must first understand the overall landscape in which information is used (Davenport 1997: 34).

Figure 5 depicts an ecological model for information management developed by Davenport. It shows the many interconnected components of the ecological approach including the information, organizational, and external environments. What follows is a brief summary of each environment based on Davenport's (1997) account. This section owes much to his book, which I quoted and paraphrased (using mainly his words) rather liberally.

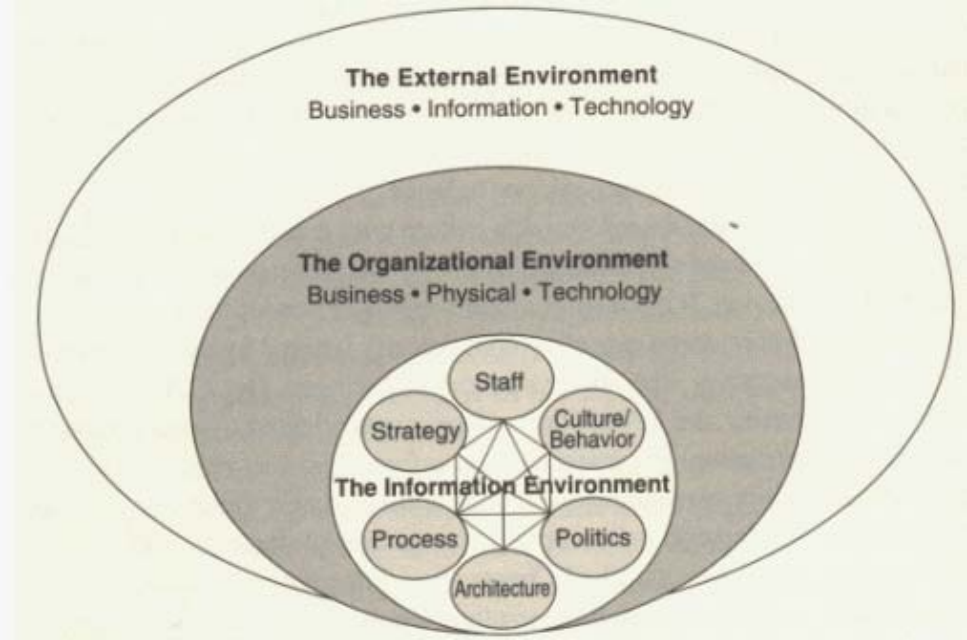


Figure 5: An Ecological Model for Information Management.
Source: Davenport (1997: 34).

8.1 The Information Environment

As indicated in Fig. 5, the information environment constitutes the core of an ecological management approach and encompasses the six most critical components of information ecology – strategy, politics, behavior/culture, staff, processes, and architecture. The information environment consists of the whole set of cross-relationships among information people, strategies and policies, processes, technology, information culture and behavior.

Information Strategy: An information strategy can potentially encompass all aspects of an information ecology. The information strategy of an organization makes its high-level "information intent" explicit in

an information-pervasive world. This strategy revolves around the question, "What do we want to do with information in this organization?" An information strategy means making choices, not carving out a master plan in stone. It should outline a set of basic goals or "principles" and be flexible.

Information strategies help organizations adapt to change and make information more meaningful for the whole organization by better allocating the information resources. Davenport (1997: 49-57) describes in detail information strategies that focus on: (1) information content (gathering, analysing and acting on the most important information, be it financial, operational, or market-oriented); (2) common information (sharing common information in order to ease communications across divisions, functions and/or business processes); (3) information processes (defining processes for collection, use and disposal of information to carry out strategic goals of the organization); and (4) new information markets (using information not just internally but selling it to outside organizations, e.g., airline industries selling flight schedules to travel agents).

Information Politics: This critical component involves the power information provides and deals with the governance responsibilities for its management, control and use. Toffler (1990) observed that in most cases the way we organize information determines the way we organize people and the vice versa. Davenport (1997: 68-72) describes four model of information governance: monarchy (an "information czar" controlling most of an organization's information), federalism (a few information elements defined and managed centrally while the rest is left up to local units), feudalism (each unit manager within an organization managing his/her information environment like lords in so many separate castles), and anarchy (every individual within the organization fending for himself or herself with no centralized approach to information management initiated by the top executives). Davenport sees these four models of information control as a continuum (Fig. 6) and emphasizes the importance of matching an organization to the political structure that best suits it.

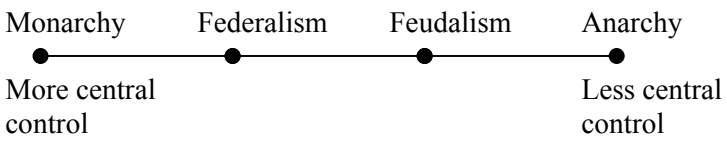


Figure 6: The Continuum of Information Control.
Source: Davenport (1997: 69).

Information Behavior and Culture: Information behavior and culture concerns with the ways in which individuals, groups or organizations deal with information and concentrates on their approaches, attitudes and behaviors towards information. It comprises of different ways of using information (such as browsing, searching, sharing, hiding, ignoring, and making use of it). Davenport considers these two related factors most important in creating a successful information environment while, at the same time, they are the toughest to change:

> All of a company's information behaviours, good or bad, make up its information culture. Particular information cultures determine how much those involved value information, share it across organizational boundaries, disclose it internally and externally, and capitalize on it in their businesses (Davenport 1997: 35).

Information Staff: Information staff consists of various professionals dealing with content (e.g., librarians, information specialists, and indexers), information technology (e.g., system designers, database administrators, network specialists, and programmers) and other information works (e.g., management accountants, business, market, or financial analysts). Davenport rightly points out that:

> People are still the best identifiers, categorizers, filterers, interpreters, and integrators of information. . . . The all-important information staff of a company handles the more valuable

forms of information, such as organizational knowledge and best practices. If the information in these categories is to be of value, it must be continually pruned, restructured, interpreted, and synthesized – all tasks that computers do poorly (Davenport 1997: 35-36).

Information Processes: This component concentrates on how information work gets done. Davenport (1997: 134-152) describes a generic information management process that has four steps: determining information requirements (identifying how managers and workers make sense of their information environments); capturing information (scanning, categorizing, formatting and packaging information); distributing information (applying a combination of "pull" and "push" technologies to draw attention to the available information sources and services); and using information (assessing the information use).

Information Architecture: This component is a simple guide to the structure and location of information within organization. It is simply a set of aids that match information needs with information resources. The information architecture can be descriptive as well as prescriptive including maps, directories and standards as well as engineered models. Davenport (1997: 158-161) draws attention to the fact that most information architectures will not change culture and behaviors of information users and information staff. They may still prefer the old way of doing things.

8.2 The Organizational Environment

No information environment exists in and of itself. It is usually a part of the broader organizational environment and has to take into account of the organization's overall business situation, existing technology investment, and physical arrangement.

An organization's overall business situation consists of its business strategy, business processes, organizational structure and culture, and human resources. The information environment cannot fully function unless it is considered as an integral part of the overall organization and involved in the creation and development of such strategies and processes.

Existing investments of an organization on information technology determines to a great extent how the information environment carries out its responsibilities. Davenport offers the following general guidelines when investing in new technologies:

- A high degree of network interconnectedness facilitates the exchange of information in organizations;
- Knowledge and information workers require personal computers or workstations on each desktop;
- Effective information management increasingly involves providing network access to internal information repositories with many CD-based databases;
- The effective management of organizational information environments increasingly demand network management software;
- An increasing number of sophisticated software packages can help manage and distribute qualitative or document-based information in organizations;
- For external information access and communications, use of the Internet is increasingly becoming a necessity; and
- For some companies, the World Wide Web can be a new means of organizing and accessing information (Davenport 1997: 184-186).

Physical arrangement in an organization concerns with "where individuals and groups are located in relation to others with whom they work. This component also consists of the physical structures – building layouts, offices, furniture – in which people work. Finally, it includes the physical appearance and

dispersal of information" (Davenport 1997: 186). Some physical arrangements facilitate information sharing while others may hinder communication of information within the organization.

8.3 The External Environment

The information ecology of any organization is affected by external factors. Among these factors are government regulations, political and cultural trends in a country and in the world, business markets (customers, suppliers, competitors, regulators and public policy), technology markets (infrastructural, current-use, and innovative technologies) and information markets (buying and selling information), and the competitors' success or failures. Such factors are beyond the control of an organization. In order for an organization to interact with the external environment, it has to:

- *adapt* to the outside world by closely monitoring government regulations and developing customer and supplier interfaces;
- *scan* that world for changes to which it must respond by identifying what external information is required, deciding where to look for it, buying and using it; and
- *mold* the outside world, through public relations and through marketing information products or services, to its own competitive advantage (Davenport 1997: 193-217).

That concludes our sketchy description of Davenport's ecological model of information management. Although Davenport acknowledges the crucial role of technology in information management (1997: 183), he places more emphasis on other components that we outlined above. He quotes Tom Peters, the management guru, as saying "Success in information management is 5% technology, and 95% psychology" (p. 175) and thinks that:

> The tools most frequently employed to design information environments derive from the fields of engineering and architecture; they rely on assumptions that may be valid when designing a building or a power generator, but rarely hold up in an organization. The sheer volume and variety of information, the multiple purposes to which it is put, and the rapid changes that take place overwhelm any rigorous attempt at central planning, design, or control (Davenport 1997: 8).

Davenport (1997: 228) believes that "No company . . . will ever achieve a true competitive advantage without adopting more human-oriented approaches to managing it. . . . It's time to look to ourselves for the information answers."

9.0 CONCLUSION

The Chinese proverb, "May you live in interesting times," appears to have already been materialized: we do live in interesting times. The famous poet Paul Valery might have thought of the same thing when he said: "The future ain't what it used to be." This is certainly the case in the provision of electronic information services. The proliferation of electronic information products and services, increasing availability of information processing, storage and communication technologies make the jobs of information managers all the more interesting. Evolving economic paradigms based on the use, rather than ownership, of electronic information sources are a challenge for electronic information managers. Libraries and information centers are no longer "the only game in town": they have to compete with other non-profit and for-profit providers of electronic information services. Information managers have to run their non-profit library and information centers as dynamic institutions and respond quickly to the user needs. They have to adapt to changes in the immediate information, organizational and externals environments and cope with business, technology, and market pressures. They have to cooperate with other entities within their organizations to develop more innovative information services involving the use

of both internal and external sources. They have to provide the best quality personalized information services as and when their users want and deliver those services to wherever their users reside. Only then can they survive, flourish, and be "better than the best."

10.0 REFERENCES

Bergman, M.K. (2000 July). *The deep Web: surfacing hidden value*. (White Paper). [Online]. Available: http://www.brightplanet.com/technology/deepweb.asp [29 June 2004]. Also appeared in: *The Journal of Electronic Publishing*, 7(1), August 2001. [Online]. Available: http://www.press.umich.edu/jep/07-01/bergman.html. [28 June 2002].

Berkhout, V. (2001). GÉANT: a giant step for European research. Presentation given at the TERENA Networking Conference 2001, 16 May 2001. [Online]. Available: http://www.dante.net/geant/presentations/vb-geant-tnc-may01/ [28 June 2002].

Brown, J.S. and Duguid, P. (2000). *The social life of information*. Boston, MA: Harvard Business School Press.

Bush, V. (1945 July). As we may think. *Atlantic Monthly*, 176(1): 101-108. [Online]. Available: http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm [29 June 2004].

Davenport, T.H. (1997). *Information ecology: mastering the information and knowledge environment.* New York: Oxford University Press.

Digital Library SunSITE Collection and Preservation Policy. (n.d.). [Online]. Available: http://sunsite.berkeley.edu/Admin/collection.html [29 June 2004].

Dumitriu, (2004 May 1). The E.U.'s definition of terrorism: The Council Framework Decision on combating terrorism, Part 1 of 2. *German Law Journal*, No. 5. [Online]. Available: http://www.germanlawjournal.com/article.php?id=434 [30 June 2004].

Dyson, G.B. (1997). *Darwin among the machines*. London: Penguin Books.

Evans, D. (1995). Meno's puzzle. In *The concept of knowledge: The Ankara Seminar.* Ed. by İ. Kuçuradi and R.S. Cohen. (pp. 97-102). Dordrecht: Kluwer.

Grady, D. (1997 July). Best bite of summer. *Self*, 19(7): 124-125.

Graham, P.S. (1994). Intellectual preservation: electronic preservation of the third kind. *The LIBER Quarterly,* 4: 163-174.

Guernsey, L. (1999, July 8). Seek—But on the Web, you might not find. *New York Times*, p. B8. [Online]. Available: http://www.greenwichacademy.org/studentlife/intranet/campus/tech_support/training/July8webarticle.htm [28 June 2002].

Hart, C.H.L. (1995). Mass customization: conceptual underpinnings, opportunities and limits. *International Journal of Service Industry Management*, 6(2): 36-45.

Hedstrom, M. (1998). Digital preservation: a time bomb for digital libraries. [Online]. Available: http://www.uky.edu/~kiernan/DL/hedstrom.html [29 June 2004].

Kahle, B. (1997 March). Preserving the Internet. *Scientific American,* 276(3): 82-83. [Online]. Available: http://www.sciam.com/0397issue/0397kahle.html [10 December 2001].

Kyrillidou, M. and Young, M. (2001). ARL statistics trends: an introduction. [Online]. Available: http://www.arl.org/stats/arlstat/00pub/intro.html [29 June 2004].

Lakoff, G. (1990). *Women, fire and dangerous things: what categories reveal about the mind.* Chicago: University of Chicago Press.

Lindley, J.A. (2000). Strategic issues in electronic librarianship. *Bilgi Dünyası*, 1(2): 330-341.

Lyman, P. & Varian, H. (2000). How much information? 2000. [Online]. Available: http://www.sims.berkeley.edu/how-much-info/index.html [29 June 2004].

Lyman, P. & Varian, H. (2003). How much information? 2003. [Online]. Available: http://www.sims.berkeley.edu/research/projects/how-much-info-2003/ [29 June 2004].

Manes, S. (1998, April 7). Time and technology threaten digital archives…but with lack and diligence treasure-troves of data can be preserved. *New York Times.* Cited in: Anne Muller, e-mail message sent to EPIC-LST@NIC.SURFNET.NL [14 August 1999].

Mok, C., Stutts, A.T. and Wong, L. (2000). Mass customization in the hospitality industry: Concepts and applications. [Online]. Available: http://www.hotel-online.com/Neo/Trends/ChiangMaiJun00/CustomizationHospitality.html [29 June 2004].

O'Neill, E.T., Lavoie, B.F. and Bennett, R. (2003 April). Trends in the evolution of the public web: 1998-2002. *D-Lib Magazine*, 9(4). [Online]. Available: http://www.dlib.org/dlib/april03/lavoie/04lavoie.html [30 June 2004].

Postman, N. (1993). *Technopoly: the surrender of culture to technology.* New York: Vintage Books.

Plato's Meno. (1971). Ed. by Malcolm Brown, tr. by W.K.C. Guthrie. Indianapolis, NY: The Bobbs-Merrill Company.

Preserving digital information. (1996). [Online]. Report of the Task Force on Archiving Digital Information commissioned by The Commission on Preservation and Access and The Research Libraries Group. May 1, 1996. Available: ftp://ftp.rlg.org/pub/archtf/final-report.pdf [29 June 2004].

Reich, R.R. (2002). *The future of success: working and living in the new economy*. New York: Vintage Books.

Rifkin, J. (2000). *The age of access: how the shift from ownership to access is transforming modern life.* London: Penguin Books.

Rosenfeld, L. and Morville, P. (1998). *Information architecture for the World Wide Web.* Sebastopol, CA: O'Reilly.

Schiesel, S. (1999, May 4). Nortel plans new product to bolster optical networks. *The New York Times,* [Online]. Available: http://www.nytimes.com/library/tech/99/05/biztech/articles/04nortel.html [29 June 2004].

Sewell, R.G. (comp.) (2001, May). Library materials budget survey 2000/2001: results of the 2000/2001 library materials budget survey of the ALCTS/CMDS/chief collection development officers of large research libraries discussion group. [Online]. Available: http://www.arl.org/scomm/lmbs/lmbs2001.html [29 June 2004].

Shapiro, C. and Varian, H.R. (1999). *Information rules: a strategic guide to the network economy*. Boston, MA: Harvard Business School Press.

Svenonius, E. (2000). *The intellectual foundations of information organization.* Cambridge, MA: MIT Press.

Toffler, A. (1970). *Future shock*. New York: Random House.

Toffler, A. (1990). *Powershift: knowledge, wealth, and violence at the edge of the 21st century*. New York: Bantam Books.

Tonta, Y. (1991). A study of indexing consistency between Library of Congress and British Library catalogers. *Library Resources & Technical Services,* 35(2): 177-185.

Tonta, Y. (2001). Collection development of electronic information resources in Turkish university libraries. *Library Collections, Acquisitions & Technical Services*, 25: 291-298.

Tonta, Y. (2003). The personalization of information services. *Information Management Report,* August 2003, pp. 1-6.

UNODC. (2004). United Nations Office on Drugs and Crime. Definitions of terrorism. [Online]. Available: http://www.unodc.org/unodc/terrorism_definitions.html [30 June 2004].

Varian, H. (1995 September). The information economy. *Scientific American,* 273: 161-162.

Waddington, P. (1997). Dying for information? A report on the effects of information overload in the UK and worldwide. [Online]. Available: http://www.cni.org/regconfs/1997/ukoln-content/repor~13.html [29 June 2004].

New Initiatives for Electronic Scholarly Publishing: Academic Information Sources on the Internet

Ana Maria Ramalho Correia
Instituto Superior de Estatística
e Gestão de Informação
Universidade Nova de Lisboa
Campus de Campolide
1070-124 Lisboa
PORTUGAL

acorreia@isegi.unl.pt

José Carlos Teixeira
Departamento de Matemática
Universidade de Coimbra
Largo D. Dinis – Apartado 3008
3001-454 Coimbra
PORTUGAL

teixeira@mat.uc.pt

ABSTRACT

This paper will trace the evolution of scholarly communication from the 17th century up to electronic journals, e-prints, e-scripts, electronic theses and dissertations and other digital collections of grey literature that emerge, with different degrees of acceptance in several disciplines, in the context of the new publishing models of the present day. The "open access or archiving/depositing" of electronic copies of scientific and scholarly research papers, theses and dissertations and other academic materials into networked servers, aims to ensure the widest possible dissemination of their contents by making them freely available on the public Internet, facilitating full use by readers without financial, legal or technical barriers, other then those related with gaining access to the Internet itself. The free and unrestricted online availability of this literature gives readers an opportunity to find and make use of relevant literature for the advancement of science and technology. From the point of view of authors, open archiving brings increased visibility and widens the readership and impact of their research work. For researchers in poorly resourced organizations and/or countries, access to this open archive material has great potential value, as it facilitates the retrieval of research results through the Internet. At the same time, it also allows scientists in these organizations/countries to participate in the development of the global knowledge base. The paper will outline the strategic factors that impact on the acceptance of these scholarly materials, available through "open access or archiving/depositing". Several international initiatives aiming at the creation of a global network of cross-searchable research materials are underway. Some of the more relevant ones will be reviewed. The impact that the availability of these novel scholarly materials is having, on changing roles and responsibilities of information managers, will also be highlighted.

1.0 INTRODUCTION

One of today's most rapidly evolving and pervasive aspects for library and information managers is the growth of published material appearing first, or only, in electronic/digital format; this is particularly true of scientific research publications. More and more authors are placing their papers directly on the Web, traditional publishers are moving into the electronic publishing business, and other companies, new to the sector, are moving into electronic publishing (Oppenheim 1997).

In a relatively short period of time (since the 1970s), electronic information resources have expanded from a few dozen computerized bibliographic databases to include the overwhelming diversity of services and products created by the electronic publishing industry. These are available in digital formats like CD-ROM and DVD, as well as through the Internet. In parallel, new publishing models, using the Internet, are being evolved, within the scholarly communication environment. The latter are being

motivated by the opportunities created by the information technologies and Internet development, in parallel with the changing economics of publishing. A strong movement, among researchers and academics (user community), seeks to free scientific information and provide unrestricted access to it for all scientists, scholars, students and the interested public.

This means that a wide range of materials is made available through the Internet and can be accessed without using the library. These include preprints (e-prints), research manuscripts (working papers, technical reports and memoranda, the so called e-scripts), Electronic Theses and Dissertations (ETDs), all freely available in electronic format.

For their part, scientific publishers, as part of their strategy to attract customers, are making freely available, high quality, current and authoritative material, *e.g.* full listings of titles, tables of contents, sample copies of journals, as well as supporting the development of e-print servers, etc. They are even creating online communities around an area of shared interest, *e.g*. *BioMedNet* (http://www.bmn.com, the website for biological medical researchers, owned by Elsevier Science), *ChemWeb* (http://www.chemweb.com/, owned by Elsevier, which is a resource covering a wide range of information for those in research chemistry, the chemical industries and related disciplines.

The availability of these novel electronic resources has altered the traditional longstanding links between information professionals and their clients (Fecko 1997:6). In the past, the library was the interface between the user and a vast amount of published and unpublished information (Oppenheim 1997: 398), which was made available in hard copy, via online databases or in CD-ROM format.

Nowadays, information managers are handling an increasing proportion of new electronic/digital materials. These are produced by the electronic publishing industry and increasingly, in the case of Special and Academic Libraries, by authors using the alternative publishing models mentioned previously.

This paper will present an overview of the evolution of scholarly communication, from the dawn of the scholarly journals up to the new electronic publishing models. It will discuss e-prints repositories, proposing a classification that illustrates the different approaches to content and its creation. The benefits of e-prints, as well as the evolving prior publication policies of scholarly journals, regarding e-prints, will be discussed.

The challenges of launching, building, maintaining and developing digital libraries of scholarly materials, including Electronic Theses and Dissertations, will also be introduced.

The paper concludes by pointing out how these publishing genres are bringing new challenges and creating new opportunities in the evolving roles of contemporary librarians and information managers.

2.0 EVOLUTION OF SCIENTIFIC COMMUNICATION

The origins of formal scholarly publishing date back to the 17th century, to the correspondence among scholars (Boyle's Invisible Colleges) in England (Meadows 1998: 5; Oppenheim 2000: 361). Groups of scholars used to meet regularly to present papers and discuss research results, under the auspices of the Royal Society. They were also corresponding by private letters, publishing short accounts of the work in progress to update those members who were unable to attend the meetings. As the volume of correspondence grew, various scholarly journals emerged as a more efficient means to exchange information in a broader sense. *Journal des Sçavans* and the *Philosophical Transactions of the Royal Society of London* were amongst the first titles to be published (Schauder 1994; Meadows 1998: 6 - 8).

The scholarly journal may have started in the 17th century as a means of communication – dissemination of important research findings to the wider research community – but throughout the 18th and 19th

centuries the "nature of journals slowly changed, resulting in a relative decline in the importance of learned society proceedings, and the successful creation of more specialised journals, reflecting the fragmentation of knowledge into more specialised disciplines" (Day 1999). The scholarly journal soon assumed the additional functions of registering "ownership" – the "scientific paternity", according to Guedon (2001) – establishing "priority" over a particular scientific discovery or advance, and of "packing" current communication into an indexed and readily accessible archive – a "public registry of scientific innovation" (Guedon 2001).

In the 19th century, yet another function was added. Publication of articles in journals came to be the prime indicator of professional standing for research professionals and the organizations that employed them (Schauder 1994: 75). Thus, while primarily allowing academics to inform peers of their findings and to be informed by them, the peer-reviewed journal also fulfilled other requirements (Boyce 2000: 404; Day 1999; Rowland 1997):

- author evaluation – providing a means for judging the competence and effectiveness of authors;
- author recognition – publication in refereed journals, raising an author's profile, improving chances of funding for future research contracts, tenure or promotion;
- validation of knowledge and quality control – through the process of peer review of submitted papers;
- historical record – maintaining the record of progress of science through the years;
- archive – providing a repository for the body of knowledge about a particular field.

The appearance and success of the electronic journal should be seen in the context of the development of electronic information systems evolving from the application of early electronic computers to scientific and technological information management.

In fact, the appearance of a large number of scientific documents produced by the Allies during World War II and the acquisition of the Axis documents following the War, triggered the need for new ways of organizing, storing and accessing this enormous body of information. Vannevar Bush, a former President of the Massachusetts Institute of Technology (MIT) and Director of the US Wartime Office of Scientific Research and Development, in his paper published in 1945, envisioned a system to store information (such as books, pictures, articles, newspapers and business correspondence) and which could be searched from a scientist's desktop, using a series of navigational links (Bush 1945). Although the Bush system – *MEMEX* – was microform based, it can be considered as the precursor to the modern hypertext systems of the Web (Large, Tedd and Hartley 1999: 43).

The expansion of research, since World War II, brought an exponential growth in the number of scientists over the years due to the increase of R&D activity, financed by industry, in parallel with that sponsored by public funds; the nature of research also evolved, over time, from specialized to interdisciplinary. All of these trends gave rise, over the decades, to different methods of scientific communication (Tenopir and King 2000: 18 - 21).

Since the 1960s, several government agencies in USA, United Kingdom, the Union of Soviet Socialist Republics and Japan, have been supporting a significant research effort aiming to find solutions to a number of problems within scientific (and technical) communication in general and scientific journals in particular. The problems addressed include the "information explosion", increasing publishing costs (and therefore prices), delays in publishing and distribution inefficiencies – the *serials crisis* (Tenopir and King 2000: 21 - 22; Large *et. al*. 1999: 43 - 44).

Many innovations have been considered and introduced, such as electronic publishing, digital processing of information and storage of large sets of data. These gave rise to the electronic journal, abstracting and

indexing journals produced electronically and the emergence of electronic databases of bibliographic information linked to the printed product (such as, *Chemical Abstracts*, *Engineering Index*, *Index Medicus*, etc…) (Large, Tedd and Hartley,.. 1999: 43).

In the 1980s, the R&D efforts to address the problems within scientific and technical communication, funded by government organizations, were terminated. New research projects, aiming to prove the feasibility of the electronic journal, were financially supported by publishers and, in the UK, by the British Library Research and Development Department (BLRDD). Several electronic journal experiments were launched over this period, for example (Tenopir and King 2000: 24):

- *ADONIS* – a journal article delivery service using the CD-ROM as the medium (a project of Elsevier, Springer and Blackwell Science, sponsored by British Library and the European Commission); and others as
- *Red Sage, BLEND, ELVYN* and *TULIP*.

The development of the Internet and then the Web, in the '90s, has had a significant impact on the decline of the traditional printed journal as the pre-eminent vehicle for scholarly communication. These enabling technologies aroused the interest of some imaginative researchers who in a vociferous and radical way sought to convince the academic community that the printed journal would disappear, within a few decades (Harnad 1990; Odlyzko 1995) (quoted in Tenopir and King 2000: 24). Despite the fact that many journal publishers had begun to set up Web-based services, to give access to electronic versions of their existing printed journals, these radical proponents of electronic communication assumed that this is just replicating, in the new medium, the *status quo* of the print version. Several models of self-publishing (sometimes called self-archiving) have been proposed, using the new enabling technologies as a means of returning the responsibility and ownership of scholarship to its creators (Okerson 1992).

One of the key assumptions behind this movement, "to develop innovative publishing models for scientific communication is that, when scholars and scientists publish in peer-reviewed journals, they are not interested in monetary reward (royalties) but in having their work read, used, built-upon and cited" (Harnad and Hemus 1998). Researchers and academics are only too aware that job opportunities, tenure, promotion and merit pay are all dependent on the attention their papers receive; consequently authors of journal articles seek impact instead of royalties (Cronin and Overfelt 1995; Walker 2002).

As a result, the established scholarly journal system has been experiencing significant challenges to its continuing pre-eminence, due to several factors. Some of these are pointed out by Sompel and Lagoze (2000) and include:

- the rapid advances in most scholarly fields means that the turnaround time of the traditional publishing model is an impediment to the speedy dissemination of R&D results among peers (Tenopir and King 2000);
- the traditional model, requiring full transfer of intellectual property rights from author to publisher, works against the promotion and wide dissemination of results and obtaining peer recognition and visibility among colleagues (Bachrach *et al*. 1998);
- the current peer review, as an essential feature of the scholarly review process, is too rigid as it stands at present and often works against the expression of new ideas, by favoring publication of papers originating from authors in the more prestigious organizations and by causing unacceptable delays in publication (Harnad 1998; 1999; 2000);
- the disparity between increases in journal subscription rates, often exceeding rates of inflation and affordable library budgets (Tenopir and King 2000; Walker 1998); the so called "serials pricing crisis", which is seriously jeopardizing the economic viability of the printed system of scholarly communication, stems from several contributing factors, as Bot and Burgemeester (1998) point out:

> General inflation and increase in size – (more pages per issue, more issues per volume, more volumes per year) – in conjunction with a dramatic decrease in personal subscriptions, which started in the 1970s. Publishers have apparently addressed this fall in revenue by increasing institutional subscription rates, thereby causing a vicious circle of cancellations and further increases in institutional rates.

This environment has encouraged the emergence of novel publishing models for formal and informal communication among scientists, based on Internet technologies for the dissemination and communication of research materials, with functionalities that far exceed those existing in print world. As well as promoting rapid access to information existing in scientific documents, in many cases without a fee, they also facilitate access to large amounts of multimedia materials on the Web and stored in databases, like biological sequences, time series, videos, etc.

The new publishing models being tested in different disciplines include, as outlined recently by Kling, Specter and McKim, (2002):

- *electronic journals* – an edited package of articles that is distributed to most of its subscribers in electronic form. Articles from an e-journal may and probably will be printed for careful reading; they might be stored in libraries in a printed form, for archival purposes. However, e-journals are accessed primarily in electronic form (Kling and McKim 1999: 891). Examples of peer reviewed, pure electronic journals are: *Florida Entomologist* (http://www.fcla.edu/FlaEnt/) and those produced by the *Entomological Society of America* (http://www.entsoc.org/pubs) (Walker 2002);
- *hybrid – paper electronic* (or the *p-e*) *journal* – (this is usually the electronic version of a paper journal) – it is a package of peer-reviewed articles available through electronic channels, but whose primary distribution channels are paper based. Examples of *p-e journals* include: *Science Online*, *Nature* (http://www.nature.com/); or the *e-p* journal *(hybrid e-p)* which is primarily distributed electronically and has a limited distribution in paper form (Kling and McKim 1999: 891);
- *author's self posting* – author's posting their articles on their Web-sites (Okerson and O'Donnell 1995);
- *field wide e-print repositories* (Ginsparg 1996; Holtkamp and Berg 2001, Brown 2001a, 2001b).

In their paper, Kling, Spector and McKim (2002) have also proposed the Guild Publishing Model (GPM) as an important scholarly communication model. This is derived from the formal research manuscript series (the *e-script* series) that are sponsored by academic departments and research institutes and which finds counterparts in the electronic environment, in the areas of economics, business, demography, higher energy physics, logic and information systems. Authors quoted the following, among others, as illustrative of the *GPM*:

- *Harvard Business School Research Manuscript Series* (Business research) (http://www.hbs.edu/dor/papers.index.html);
- *The University of Western Ontario Discussion Paper Series* (Demography) (http://www.ssc.uwo.ca/sociology/popstudies/dp.html).

Some of these new electronic publishing models – based on self/open-archiving (*e.g.* deposit of a digital documents in a publicly accessible website) – have been tested by scholars, in several disciplines and are sponsored by academic departments or research institutions in response to the rising costs of materials produced by the traditional publishing industry.

There is a strong international movement that, at least in some scientific areas, seeks to make research papers available by this method. Academics and researchers, worldwide, visit them, as the first place to

look, before deciding to obtain the original document. This is especially useful for countries and/or organizations where financial restrictions prevent access to a wide range of commercially published journals.

*SPARC Open Access Newsletter (*SOAN) (http://www.arl.org/sparc/soa/), launched in July 2003 to continue Peter Suber's *Free Online Scholarship (FOS) Newsletter* (March 2001-September 2002) (http://www.earlham.edu/~peters/fos/index.htm), is a highly useful resource for keeping up to date with developments in all areas related to electronic scholarly publishing.

The frequently updated *Scholarly Electronic Publishing Bibliography* (http://info.lib.uh.edu/sepb/sepb.html), published by Charles Bailey (1992/2004), since 1992, includes two sections with relevant articles, one on *New Publishing Models* and other on *Repositories, e-prints and OAI.*

Discussion lists about issues surrounding the freeing of scientific literature are available at:

- *American Scientist open access forum* (http://amsci-forum.amsci.org/archives/september98-forum.html);
- *Nature's Forum on future e-access to the primary literature* (http://www.nature.com/nature/debates/e-access/),

where the points of view of different stakeholders, in the movement to develop innovative publishing models for scholarly materials, can be appreciated by following the debates taking place electronically.

3.0 E-PRINTS

3.1 Introduction

As discussed above, the advent of the Internet enabled researchers and academics to recognize that the information and communication technologies gave them efficient ways to share results, to combat the rise in journal costs fast outpacing a library's ability to afford them (serials crisis), to overcome the barriers raised by the full transfer of Intellectual Property Rights from author to publisher and to improve on the hitherto slow turnaround of traditional publishing. While several of their initiatives began as *ad hoc* vehicles for dissemination of preliminary results, a number of them have evolved into a more formal means for the efficient sharing of research results among peers in the field (Correia and Neto 2002).

The term *e-print* encapsulates a wide range of meanings. Originally it was defined as an electronic preprint circulated among colleagues and field specialists to obtain feedback; the concept of *e-print* was then generalised to include any electronic version of academic research manuscripts circulated by the author outside of the traditional scientific publishing environment. They may be journal articles, conference papers, book chapters or any other form of research output (Luce 2001).

An "*e-print archive*" is simply an online repository of these materials, which is publicly accessible. Some *e-prints* may be peer reviewed before being posted on the servers; others are posted without peer review and authors request feedback on the results submitted (Garner, Horwood and Sullivan 2001: 250).

Some authors prefer to make the distinctions between the two forms, using the terms "*pre-prints*", for papers before they have been refereed and unpublished papers which are usually submitted for publication, or "*post-prints*", for final peer-reviewed and published versions of papers (as such they incorporate any changes or corrections necessary to ensure publication). In some cases, the e-prints are loaded onto servers, but are also submitted to traditional journals that have a peer review process. This procedure varies depending on the discipline (*e.g.* physicists, mathematicians, computer scientists and astronomers *vs*. bio-medicine and chemistry) (Kling, Spector and McKim 2002: 2; Brown 2001a: 188)

and on the prior publication policies of some journals towards publication and subsequent publication of e-prints (Brown 2001a:188), as will be discussed later in this paper.

Typically, an e-print archive is normally made freely available on the web, with the aim of ensuring the widest possible dissemination of its contents, to inform colleagues about research in progress and to seek expert comment (Pinfield, Gardner and MacColl. 2002).

The first *e-print* server was the *Los Alamos Physics Archive*, presently known as *arXiv.org*, which was created in 1991 by Ginsparg (Ginsparg 1996; Luce 2001; McKiernan 2000) at the Los Alamos National Laboratory, to give access to pre-prints in the domain of high-energy physics. Since July 2001, this archive has been located at Cornell University. During the last decade, it has evolved to become the primary means of scholarly communication and the largest non-peer review research works deposit available, worldwide. It is a fully automated electronic archive for research papers in physics and related disciplines, mathematics, non-linear sciences and computational linguistics. The *arXiv.org* archive is mirrored in more than a dozen countries.

The success of this new method of scientific communication and its acceptance among researchers and academics can be appreciated from the following data:

- In the beginning of 1999, *arXiv.org* contained 100,000 articles and the number of articles downloaded, yearly, exceeded 7 million, indicating that each article is downloaded at least 70 times, on average (this value refers only to the *Los Alamos* server and excludes the mirrors) (Odlyzko 2002);
- Two years later (2001), the number of articles made available by *arXiv.org* was some 150,000 and is growing at about 30,000 papers per year (Harnad 2001b).

Additionally, those in the domains of management, business and finance circulate “working papers” in a similar way to the physics preprints, which led to the creation of the *RePEC – Research Papers in Economics* service. Even allowing for existing differences between disciplines, in relation to the degree of adoption of electronic communication, as Kling and McKim (2000) have described, there are already, in most disciplines, examples of services providing access to electronic research papers. The following paragraph attempts to characterize the several approaches to create such services.

3.2 Classification of e-prints Collections

This is an attempt to illustrate the diversity of approaches towards *e-prints* archives existing at present, the scope of the content and the tools to facilitate access to e-prints repositories and other grey literature available in any given scientific domain. The approach used here is to provide examples, which illustrate the broad groupings:

a) Subject/discipline specific and field wide *e-prints* repositories (*e.g. arXiv* (http://www.arxiv.org), firstly in high-energy physics, latterly extended to the whole of physics, mathematics and computer sciences; *CogPrints* (http://cogprints.soton.ac.uk/), covering cognitive sciences (namely, psychology, neuroscience, linguistics, and biology); *RePEc – Research Papers in Economics* (http://repec.org), in the field of economics (Krichel 2000; Cruz and Krichel 2002); *AgEcon Search: Research in Agricultural and Applied Economics* (http://agecon.lib.umn.edu), in the field of agriculture economics, including sub disciplines such as agribusiness, food supply, natural resources economics, environmental economics, etc. (Letnes and Kelly 2002); *The Mathematics Preprint Server* (http://www.mathpreprints.com/math/Preprint/show/index.htt).

b) Central *vs.* Distributed institutional e-prints repositories

E-prints repositories are either,

- Centralised subject-based, single e-prints repositories based at single institutions [*e.g. arXiv*, at Cornell University; *CogPrints*, at Southampton University), where authors from any institution are required to submit their papers to archive, remotely by email or using a self-archiving procedure online (Pinfield, Gardner and MacColl. 2002), or
- Distributed institutional e-prints archives – whose aim is to offer academics and researchers of faculties and participating organizations, a central location for depositing their pre-publication scholarship, *e.g.*
  - *Nottingham ePrints* (http://www-db.library.nottingham.ac.uk/eprints/), University of Nottingham (Pinfield *et. al*. 2002);
  - *Glasgow ePrints Service* (http://eprints.gla.ac.uk/), University of Glasgow;
  - *University of California eScholarship Repositories* (http://repositories.cdlib.org/), University of California;
  - *The Australian National University Electronic Pre and Post Print Repository*, (http://eprints.anu.edu.au/index.html), The Australian National University.

Centralised subject-based e-prints repositories have only been taken up by a limited number of subject communities. In their turn, the institutionalised models have an interest for those institutions that decide to promote themselves in the scientific and academic community.

c) Acceptance (or not) of non-refereed papers

There are *e-prints* repositories which accept deposit of non-refereed papers without any peer review process and in many cases invite commentary from peers, together with the final peer reviewed version (*e.g. ArXiv*; *CogPrints, Netprints, Glasgow ePrints Service*). Others are repositories which only accept refereed papers (*e.g. PubMedCentral*, *BiomedCentral*, *e-BioSci*). The latter are particularly relevant in the areas of medicine.

d) Infrastructures to make research output electronically (which can be considered a special case of distributed institutional e-prints repositories) for academic publications of faculties and participating organizations, *e.g.*

- *ARNO* (NL) (http://www.uba.uva.nl/en/projects/arno), University of Twente, University of Amsterdam and Tilburg University (Bentum *et. al*. 2001); as Prinsen (2001) states,

  it aims to build an infrastructure to make research output electronically available based on distributed archives that are interlinked by subject, connected with the existing national library infrastructure, linked with the production process of publishers, and linked with digital learning environments.

e) Portals to facilitate access to domain specific grey literature, including e-prints repositories, individual and institutional web pages:

- *PhysNet – The Physics Departments & Documents Network* http://physnet.uni-oldenburg.de/PhysNet/ (Severins *et al.* 2000);
- *MATH NET – Internet Information Services for Mathematicians* (http://www.math-net.de/)
- *MareNet – Marine Research Institutions and Documents Worldwide* (http://marenet.uni-oldenburg.de/MareNet/)

f) e-prints repositories produced by commercial publishers, in response to e-prints initiatives by researchers; among others,

- *CPS – Chemistry PrePrint Server* (http://www.chemweb.com/preprint), created by the *ChemWeb.com* (the online community for chemists operated by *ChemWeb*, a subsidiary of *Elsevier Science*), making available preprints and conference proceedings (pre-conference abstracts and post conference reports) in all areas of chemistry (Wilkinson 2000; Warr 2001). *Chemistry PrePrint Server* stopped accepting submissions as of 24 May 2004.
- *ERA – Electronic Research Archive* (http://www.thelancet.com/era), *The Lancet,* free electronic research archive in international health, for "electronic publication of unreviewed and expanded papers publish an open comment on these papers" (McConnell and Horton s/d). This archive allows medical researchers to deposit papers addressing health issues of relevance to many developing countries (Chan and Kirsop 2001);
- *NetPrints*: *Clinical Medicine & Health Research* (http://clinmed.netprints.org/home.dtl), sponsored by the *British Medical Journal* (Delamothe *et al.* 1999).

3.3 Benefits Provided by the Publication in *e-prints* Servers

For scholars and academics, there are several benefits to be gained from archiving their scientific work in e-print repositories. The following are highlighted from the point of view of the researcher, as contributor/reader of the literature:

- lowering impact barriers and increasing visibility – papers become freely available for others to consult and cite (Pinfield, Gardner and MacColl 2002). Lawrence (2001 a; 2001b) and Goodrum *et al*. (2001) provide evidence that work that is freely available is more cited;
- rapid dissemination of information to a wider audience – depending on what document types are accepted in the archive (pre-prints or post-prints), online repositories help to accelerate dissemination of the research findings (Pinfield, Gardner and MacColl 2002);
- better quality and improved efficiency in the R&D activity (by avoiding duplication) and faster communication, between academia and industry (Warr 2001);
- improved archiving of scientific data – regarding this aspect, Internet technologies offers advantages of the multimedia and the supporting files, as Garner, Howard and Sullivan (2001: 252) point out,

  (…) They have the potential to improve the way the results are portrayed by including large data sets such as digitised images and the results of failed experiments (…) It is beneficial to have systems that can help scholars and scientists to learn from each other's experiences, both successes and failures. (Garner *et. al*. 2001: 252).

The e-prints offer substantially more features than their print equivalents: for example, in some cases, annotation facilities are provided to allow commentaries by peers to be posted.

As referred to above, from the institutional point of view, institutions benefit by ensuring that their research output is widely disseminated; it helps to enhance their reputation, to attract high quality researchers and obtain further research funds.

The benefits referred to are of great importance to all scientists, both as readers and contributors and to research institutions (Pinfield, Gardner and MacColl 2002).

Furthermore, the e-prints repositories bring added benefits for scientists in poorly resourced organizations or countries. By accessing e-prints repositories available anywhere in the world, they are provided with access to the global knowledge base. Equally important are the opportunities created by the *e-prints* servers which offer the possibility, for scientists in less resourced countries or organizations, to distribute local research in a highly visible way and without the difficulties and bias associated with publishing in

traditional journals, which tend to favour the publication of papers from well known authors or from known organizations in more developed countries (Chan and Kirsop 2001).

3.4 Prior Publication Policies

Currently, policies differ between journal publishers regarding the acceptance of preprinted articles for publication in their journals. For some publishers, the posting of an article in an e-Prints service is considered as prior publication, whilst for others, it is not.

Even *ChemWeb* (http://www.chemweb.com), the largest online chemical community in the world, owned by the commercial publisher *Elsevier,* is encouraging all publishers to reach a firm decision on this issue (Warr 2001:1).

For its part, *NetPrints* (http://clinmed.netprints.org/home.dtl), the preprint server sponsored by *BMJ – British Medical Journal*, which is a repository of non-peer reviewed original research within the scope of clinical medicine and health, supplies on its website lists of journals in the domain of clinical medicine, (http://clinmed.netprints.org/misc/policies.shtml) that,

- will accept submissions that have appeared on preprint servers (*e.g.* BMJ, Breast Cancer Research, Genome Biology, Journal of Biological Chemistry, the Lancet, Nature, etc…) and
- will not accept submissions that have appeared on preprint servers (*e.g.* American Journal of Clinical Nutrition, Biology of Reproduction, JAMA, Science, etc…).

In turn, Brown (2001a; 2001b) has investigated how e-prints are cited, used and accepted in the literature of physics and astronomy, and examined the philosophies of approximately 50 top-tier physics and astronomy journals regarding e-prints. This author concludes that, "e-prints have evolved into a valid facet of the physics and astronomy literature" and that although the degree of acceptance stated by journals' editors and policies, as given in journals' instructions to authors, concerning the citing of e-prints, were found to differ, the research showed that the citation rate of e-prints is increasing. (Brown, 2001b: 187, 197-198).

Generally speaking, across the disciplines, there are publishers who refuse to consider manuscripts that have been posted on web archives. These include *American Chemical Society* and *New England Journal of Medicine* and *Science*, while others, within the same discipline, do not refuse to publish papers that have appeared on e-prints repositories – like the *Royal Society of Chemistry*, the *Journal of Neuroscience* (Garner, Horwood and Sullivan 2001: 253).

3.5 Issues and Misconceptions

Some misconceptions about the newly established e-prints repositories are still hindering support for these initiatives. These misconceptions may be grouped around the following issues:

i) Fears that *e-prints* may give rise to some sort of vanity publishing and consequently have an adverse effect on research quality

   Since any individual can publish material on the Internet, there is a concern that low quality materials will appear on the e-prints archives giving rise to some "vanity press" that has not undergone the normal quality control procedures (Chan and Kirsop 2001).

   Pinfield and other authors have, however, disagreed with this point of view, by harking back to the experience of more then ten years, with the *arXiv* repository; this indicates that researchers are always concerned with their reputation and professional credibility and that quality of research has not been at stake. Publication in a preprint server is a good way to expose the paper to widespread scrutiny (Pinfield 2001a).

Furthermore, some *e-prints* servers have implemented facilities that enable the user to exercise clear options for selectively retrieving material (*OpCit*), to discuss and rank the articles, access the most recent, the most viewed, the most discussed and the highest ranked articles [as in *Chemistry Preprint Server* (Warr 2001:2)]. These facilities enhance the potential for critical scrutiny and clearly separate vanity publishing from the rest. Even so, many authors prefer to post their papers on e-prints repositories only after they have been through the refereeing process, as they do not want to jeopardise prospects for grants and promotions or the possibility of rejection by those journals of high standing, which may still have restrictive policies regarding e-print publication. This is particularly true of the medical community, where some professionals have even gone so far as to say that pre-refereed material may be dangerous in their field, if used as a basis for clinical practice.

ii) Intellectual Property Rights (IPR), particularly copyright

The issues of IPR and copyright are complex. There is uncertainty as to the ownership of research copyright and the debate is still ongoing. In most Higher Education Institutions, accepted custom and practice is that academic authors are permitted to claim and dispose of the copyright themselves. The problem is that commercial publishers of many research journals require the authors to assign copyright to the publisher before publication. A movement taking shape within the e-print community is that authors should be encouraged to retain their IPR by submitting to journals that do not require signing over the copyright or will agree to the author distributing the papers through e-prints repositories (e-distribution rights). (Pinfield, Gardner and MacColl, 2002).

Summing up, the message to authors is loud and clear; they do not necessarily need to stop submitting their work to high-profile traditional journals. They should continue to do so but at the same time retain their copyright, to enable them to make their work available in an e-prints archive.

3.6 International Initiatives towards the Creation of a Global Network of Cross-Searchable Research Materials Archives

3.6.1 OAI -Open Archives Initiative

In order to fully exploit the expansion in the number of e-prints repositories distributed across the Internet and to develop a global network of cross-searchable scholarly and research information sources, it was soon recognised that there was a need to ensure that searches could be made across different e-prints archives.

The *OAI – Open Archives Initiative* (http://www.openarchives.org/) addresses this issue; the initiative emerged from the Santa Fe Convention held in 1999. The *OAI* aims to create cross-searchable databases of research papers and make them freely available on the web by developing and promoting interoperability standards that will facilitate the efficient dissemination of content (Pinfield 2001b). Using these standards, institutions can put content on the Internet in a manner that makes individual repositories interoperable.

At the centre of this work is the *OAI Metadata Harvesting Protocol* (http://www.openarchives.org/OAI/openarchivesprotocol.htm). This creates the potential for interoperability between e-prints archives by enabling metadata from a number of archives to be harvested and collected together in a searchable database. The metadata harvested is in the Dublin Core format and normally includes information such as author, title, subject, abstract, and date (Pinfield, Gardner and MacColl 2002).

The *eprints.org* (http://www.eprints.org/), at the University of Southampton, provides free software that enables any institution to install OAI-complaint archives (i.e. using the OAI metadata tags). It is designed to run centralised, discipline-based as well as distributed, institution-based archives of scholarly publications (Chan and Kirsop 2001).

OAI-compliant e-prints servers provide value-added facilities. They can compile statistics which show authors how many times their papers have been accessed; they can also produce an online publications list by author or by academic department. Furthermore, developing services such as the *OpCit – The Open Citation Project* (http://opcit.eprints.org) (a project funded by the *Joint NSF-JISC International Digital Libraries Programme* – http://www.dli2.nsf.gov/internationalprojects/intlprojects.html), aim to provide integration and navigation through citation linking; these are value-added services in e-prints repositories, as they provide enhanced reference linking and give authors citation and impact analysis of their work (Nottingham ePrints, *About Nottingham ePrints*. Value added services).

3.6.2 BOAI – Budapest Open Access Initiative

A meeting was convened in Budapest by the Open Society Institute (OSI), on December 1-2, 2001. Its aim was to accelerate progress in the international efforts to make scientific and scholarly research results freely available on the Internet.

The participants represented many viewpoints, academic disciplines and nations. They brought first-hand experience of many of the ongoing initiatives that make up the open access movement. They discussed how best the separate initiatives could be coordinated to achieve better progress. They examined the most effective and affordable strategies for serving the interests of research, researchers, and the institutions and societies that support research. Finally, they explored how OSI and other foundations could use their resources most productively, to aid the transition to open access and to make open access publishing economically self-sustaining.

As a result, the *BOAI-Budapest Open Access Initiative* (OSI) was announced (http://www.soros.org/openaccess/) which is a "statement of principle, a statement of strategy, and a statement of commitment".

In the first instance, BOAI addresses, specifically, peer-reviewed journal articles and preprints; nevertheless, it can be extended quite naturally to

> all content for which authors do not expect payment, namely scholarly monographs on specialized topics, conference proceedings, theses and dissertations, government reports and statutes and judicial opinion. (*BOAI public statement*, in BOAI FAQ 2002 http://www.earlham.edu/~peters/fos/boaifaq.htm).

In an article posted at *Nature's Forum on future e – access*, Butler points out that:

> Research institutions and funding agencies that sign up to the *BOAI* commit themselves to making policy changes, such as creating local open-access electronic repositories, and making it compulsory for grant recipients to deposit their papers there. Individual signatories agree to deposit their research in freely available electronic repositories, and to support alternative journals as authors, editors and referees (Butler 2002).

So far (May 2002), there have been ca. 2500 individuals and ca. 150 organizations signatories to the *BOAI – Budapest Open Access Initiative* (http://www.soros.org/openaccess/view.cfm).

3.7 Roles for Librarians Regarding Self-Publishing by Researchers

The creation of e-prints archives is a response to a number of structural problems in the academic publishing industry. The library and information services at universities or research organisations should also be the natural place for e-prints services.

However, this development does not take place in isolation. Librarians must be involved. Strong alliances should be forged between librarians, scholars, scientists and researchers and those that have the

responsibility for the development of the infrastructure. In this tripartite alliance, activities for librarians and information managers should be:

- supporting users (scholars, scientists and researchers) to e-publish their materials, to facilitate self publishing and to smooth the path for potential contributors;
- exploring the document types that e-prints services can accept and that may be relevant for the nature of research in their organisations;
- providing advice to those who wish to post their documents on the e-prints services with regard to the evolution of prior publication policies of journal publishers;
- increasing awareness of the possibilities and facilities provided by e-print archives;
- persuading institutional managers and policy makers of the benefits to be gained by creation of e-prints services in their organizations.

These are major challenges but they represent the future for librarians and information managers.

4.0 ELECTRONIC THESES AND DISSERTATIONS

Theses and dissertations assume a central role in scholarly communication, as they are sometimes the only tangible deliverables, after long and expensive periods of research. As such, they are a major source of new knowledge and contain valuable research results which, when published, are extremely useful to other groups working in the same field (Gonçalves *et. al*. 2001).

The creation of digital libraries of theses and dissertations generates an environment, which significantly increases the availability of students' research for scholars and empowers universities to unlock their information resources. This environment gives rise to a number of beneficial activities (Fox 1999; Moxley 2001: 61; Suleman *et al.* 2001b):

- Improving graduate education – where universities require Electronic Theses and Dissertations (ETDs) for graduation they inspire and instigate faculty and graduate students to experiment with new mentoring models; in the past, few (printed) theses and dissertations were read, beyond the evaluation committee. Works, archived in Digital Libraries, are read by thousands, potentially millions of people worldwide;
- Empowering students to convey a richer message through the use of multimedia and hypermedia tools, animation and interactive features;
- Endowing graduates with new capabilities and eliciting the preparation of the next generation of scholars as effective knowledge workers, by providing opportunities for students to produce electronic documents, training future graduates in the emerging forms of digital publishing and information access;
- Lowering the costs of submitting and handling theses and dissertations (eliminating binding costs and shelf space);
- Increasing accessibility, visibility and readership of students' work and at the same time, this exposure attracts to the University promoting ETD, the most innovative future candidates; those who are keen to have their theses and dissertations read are often likely to obtain the best professional offers;
- Helping universities to build their information infrastructure and extend and advance digital library impact.

The *Networked Digital Library of Theses and Dissertations – NDLTD* – is an on going project which was conceived in 1987 and realized, in part, in 1997 through the efforts of Virginia Tech's Ed Fox, Gail McMillan and John Eaton (Moxley 2001: 62).

It aims to develop a federation of digital libraries, providing free access to graduate students' theses and dissertations and is a collaborative effort of universities around the world, which promotes creating, archiving, distributing and accessing ETDs (Suleman *et. al*. 2001a; 2001b).

One of the main objectives of the federation is to "provide transparent and integrated services that span NDLTD members' ETD collections, while keeping with the general desire, at each site to maintain their individual collections" (Gonçalves and Fox 2001:15).

NDLTD aims to help coordinate the international efforts related to electronic theses and dissertations; at present, it embraces thousands of students, faculty and staff at hundreds of universities on a global scale, as well as numerous companies, government agencies and other organizations (Suleman *et al.*, 2001b).

By the end of 2001, the number of members was over 120, varying from a wide range of individual universities, libraries and consortia at the state (OhioLINK), regional (Catalunya), and national (Australia, Germany, India, Portugal, South Africa) levels (Gonçalves and Fox 2001:15).

The growth of NDLTD as a truly international consortium has been fuelled by some political decisions, which were taken recently. For example, several countries – like Australia, Germany, France and India – are implementing policies at national level to guide and standardize the development of local ETD initiatives. These initiatives have been taking different forms, like the French Minister of Education who, in 2001, distributed a public letter to every university president and graduate school announcing his desire to implement ETD's at the national level; in Germany, the Conference on University Rectors distributed a similar statement (Moxley 2001: 62).

A guide to help those interested in the ETD initiative – *The Guide for Electronic Theses and Dissertations*, funded in part by UNESCO, is available online (http://etdguide.org/) as a "living document" written by ETD scholars throughout the world (Moxley *et. al*. 2001/2). It is being updated regularly and will be translated into many languages. It is an important reference work for all those interested in research and e-publishing.

Current NDLTD research developments are focusing on the creation of a union database that will provide a means to search and retrieve ETDs from the combined collections of *NDLTD* member institutions (Suleman 2001a). In order to bridge the gap between existing distributed institutional archives and a unified collection of ETDs, a metadata standard especially suited to ETDs was developed. Additional research efforts include:

> advanced search mechanisms, semantic interoperability, the design and development of multi- and cross-lingual search systems, authors' files and software modules that support the development of higher-level services to aid researchers in seeking relevant ETDs (Suleman *et. al.* 2001b).

There is no cost for institutions interested in joining the NDLTD. It is sufficient to send a letter (see: http://www.ndltd.org/join/) to the NDLTD indicating that intention. Joining only requires agreement with the goals and objectives of *NDLTD*. For members to fully benefit from services provided, it is essential to follow the standards developed by NDLTD and participate in the Union Catalogue of ETDs (http://hercules.vtls.com/cgi-bin/ndltd/chameleon) and make their content available through the NDLTD library: (http://www.theses.org).

To increase the potential benefits of a University ETD project, and enable graduate students to create richer theses and dissertations, there is a need to provide training workshops for faculty and students to complement their knowledge on word processing tools, use of templates, hyper-linking, image insertion, etc.

Ultimately, within the ETD environment, creative researchers will challenge the traditional concept of academic writing. What the ETDs are proving is that linear text is giving way to hyper textual writing, streaming multimedia, interactive chat spaces, three-dimensional modeling and features we can't even imagine right now (Moxley, 2001: 63). This will bring about a true (r)evolution in scholarly communication, as we know it today.

4.1 Networked University Digital Library

A related project is the *Networked University Digital Library NUDL,* a worldwide initiative that again is taking advantage of digital library technology. The NUDL looks beyond ETDs and focuses on the challenges of launching, building, maintaining and developing a digital library for the totality of scholarly materials, including ETDs (http://www.nudl.org). It addresses the task of making the intellectual capital produced in universities, around the world, "more accessible, stimulating technology transfer, international collaboration and knowledge sharing across all disciplines" (Gonçalves and Fox, 2001: 14).

NUDL is focused on making available the work of scholars, writing in diverse languages, for wider use; it seeks to improve the sharing of knowledge and availability of university works, providing electronic storage and preservation (*OpCit*. 2001: 14).

NUDL builds upon the foundations of the NDLTD and expands its objectives by supporting graduate students, other university related activities and materials and specialised services for different disciplines and communities. At present, the requirements of two communities are receiving special attention: Computing and Physics (*OpCit*. 16-7).

4.2 Roles for Librarians and Information Managers Regarding ETDS

Academic librarians and information managers have a very important future role to perform within their organizations, regarding ETDs and other intellectual material produced by academics and scholars.

The following services provide access on a commercial basis to Theses and Dissertations:

- *UMI's Dissertation Services* (http://www.umi.com/hp/Support/DExplorer/)
- *Dissertation.com* (http://dissertation.com/)
- *Diplomica.com* (http://www.diplomica.com/)

Academic librarians and information managers, responsible for access, use and preservation of information available in the Higher Education institutions, should also be active promoters of digital scholarship. To this end, they can collaborate with the Computing departments to:

- advocate to the Boards/Presidents/Rectors the potential of embracing ETDs and the advantages of being members of NDLTD;
- facilitate training to meet faculty and student needs as authors and as supervisors in the digital environment;
- study and propose solutions regarding archiving and preservation of the evolving ETD genre;
- complete the metadata provided by ETD authors in order to promote efficient retrieval.

Librarians and information managers should also be the champions of promoting University collaboration at state, national and international level, contributing to the definition and establishment of open, internationally accepted standards that would increase access to the wealth of scholarly information existing in ETDs. They must be aware of these developments in Electronic Scholarly Publishing and prepare their user community to take full advantage of the benefits to be gained.

5.0 OTHER DIGITAL COLLECTIONS OF GREY LITERATURE – DIGITAL LIBRARIES OF S&T RESEARCH REPORTS

The birth of report literature, one of the first types of grey literature, is associated with the development of aeronautics, the sector in which the first set of reports was produced. *Reports & Memoranda* were published by the UK's *Advisory Committee for Aeronautics* in 1909 and the first report published in the USA was in 1915, by the *National Advisory Committee of Aeronautics* (NACA) (Luzi 2000).

With the build-up of scientific research to support the war effort, in the1940s, the use of report literature to exchange information expanded greatly in countries such as Germany, UK and USA. As every researcher, practitioner and academic knows, a considerable amount of valuable scientific information is being produced at all levels of government, academia, business and industry, in print and electronic formats, outside the normal channels of publication and distribution. It is often the case that security, privacy or confidentiality arrangements make identification and access to these resources difficult, if not impossible. Diluted and delayed accounts may appear in journal articles or in books but in many cases, the original report or account paper is the only source of the research results (Correia and Borbinha 2001; Correia and Neto 2002).

Over the past few years there has been a political will to encourage cross-sectoral collaboration in the drive for a knowledge society and to increase productivity. Alongside this political background, the explosion in the use of Internet is facilitating a common and accessible information and communication tool (Needham 2001:30). In this environment, digital libraries of research reports assume an important place in the context of engineering communication and several projects are in place to promote the efficient circulation of this information, including following:

- *MAGiC 2 – Developing the UK National Reports Catalogue*, (http://www.magic.ac.uk), a project aiming to provide UK engineering community with a greater awareness of and access to key collections of technical reports (Needham 2001);
- *National Advisory Committee for Aeronautics (NACA) Technical Report Server*, NASA, USA (http://naca.larc.nasa.gov/) (Nelson 1999);
- *Networked Computer Science Technical Reference Library* (http://www.ncstrl.org) – provides unified access to technical reports and e-prints from computer science departments, institutes and laboratories. (Davis and Lagoze 2000);
- *Caltech Computer Science Technical Reports* (http://caltechcstr.library.caltech.edu/information.html).

6.0 CONCLUSION

There are unmistakable signs of a (r)evolution in scholarly communication – scientists, researchers, academics and librarians are taking charge to produce information products and services over which they retain control, maintaining their independence from commercial, for-profit publishers. This implies that the information professional has to become skilled in the use of new applications, including:

- creation and management of collections integrating resources in a variety of formats;
- establishing links between library catalogs and the new electronic scholarly materials available on the Internet;
- address new issues created by the long term preservation of these new scholarly materials;
- increasing scientists' awareness of these new sources; this is increasingly relevant at a time when the budgets allocated to buy external scientific information is reducing for most S&T worldwide;

- supporting potential authors by providing training on electronic publishing;
- enhance user-created metadata.

Librarians and information managers should be looking for the opportunities, created by the availability of these novel scholarly materials, to play a significant part in the creation of new knowledge.

7.0 REFERENCES

American Scientist open access forum. [Online]. Available: http://amsci-forum.amsci.org/archives/september98-forum.html [1 July 2004].

Bachrach, S. *et. al.* (1998). Intellectual property: who should own scientific papers, *Science* 281 (5382):1459-60. [Online]. Available: http://www.sciencemag.org/cgi/content/full/281/5382/1459 [4 June 2002].

Bailey, C. (2004). *Scholarly electronic publishing bibliography* (version 53, 12 May). [Online]. Available: http://info.lib.uh.edu/sepb/sepb.html [1 July 2004]. *New publishing models*. [Online]. Available: http://info.lib.uh.edu/sepb/models.htm *Repositories, e-prints and OAI*. [Online]. Available: http://info.lib.uh.edu/sepb/techrep.htm

Bentum, M., Brandsma, R., Place, T. and Roes, H. (2001). Reclaiming academic output through university archive servers. *The new review of information networking*, 7: 251 - 264. [Online]. Available: http://drcwww.kub.nl/~roes/articles/arno_art.htm [1 July 2004].

Bot, M. and Burgemeester, J. (1998). *Costing model for publishing an electronic journal*, Utrecht: PricewaterhouseCoopers. [Online]. Available: http://cwis.kub.nl/~dbi/users/roes/erclaw/ejclwp6a.pdf [4 June 2002].

Boyce, P. (2000). For better or for worse. Preprints are here to stay. *College and Research Libraries News,* 61 (5): 404 - 407 (414).

Brown, C. (2001a). The Coming of age of e-prints in the literature of physics. *Issues in Science and Technology Librarianship*, Summer 2001. [Online]. Available: http://www.library.ucsb.edu/istl/01-summer/refereed.html [1 July 2004].

Brown, C. (2001b). The e-volution of preprints in the scholarly communication of physicists and astronomers. *Journal of the American Society for Information Science and Technology* 52 (3): 187 - 200.

Brown, C.M. (1999). Information seeking behaviour of scientists in the electronic information age: astronomers, chemists, mathematicians, and physicists. *Journal of the American Society for Information Science* 50 (1): 929 - 943.

Budapest Open Access Initiative: Frequently Asked Questions. [Online]. Available: http://www.earlham.edu/~peters/fos/boaifaq.htm [1 July 2004].

Bush, V. (1996/1945). As we may think. *Interactions*, 3 (2): 35 - 46. Originally published in *Atlantic Monthly* 176 (1): 101 - 108. [Online]. Available: http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm [1 July 2004].

Butler, D. (2002). Soros offers open access to science papers. *Nature Web Debates*, 14 February 2002. [Online]. Available: http://www.nature.com/nature/debates/e-access/ [4 June 2002].

Chan, L. and Kirsop, B. (2001). Open archiving opportunities for developing countries: towards equitable distribution of global knowledge. *Ariadne* 30. [Online]. Available: http://www.ariadne.ac.uk/issue30/oai-chan/intro.html [1 July 2004].

Correia, A.M.R. and Borbinha, J. (2001). Deposit scientific and technical gray literature: a case study. In ISAÍAS, P. (ed.). *New developments in digital libraries: Proceedings of the 1st International Workshop on New Developments in Digital Libraries, NDDL 2001, in conjunction with ICEIS 2001 – 3rd International Conference on Enterprise Information Systems.* Setúbal, July 2001, 8 - 19.

Correia, A.M.R. and Neto, M.C. (2002). The role of *ePrint* archives in the access to and dissemination of scientific gray literature: *LIZA* – A case study by the National Library of Portugal. *Journal of Information Science* 28 (3): 231 - 242.

Cronin, B. and Overfelt, K. (1995). E-journals and tenure. *Journal of the American Society for Information Science* 46 (9): 700 - 703.

Cruz, J. M. and Krichel, T. (2002). Automated extraction of citation data in a distributed digital library. Isaías, P. (ed.) *New developments in Digital libraries. Proceedings of the 2nd workshop on New developments in Digital libraries (NDDL 2002)* in conjunction with *ICEIS 2001 – 4th International Conference on Enterprise Information Systems*, Universidade de Castilla e La Mancha, April 2002.

Davis, J. and Lagoze, C. (2000). NCSTRL: Design and deployment of a globally distributed digital library. *Journal of the American Society for Information Science* 51 (3): 273 - 280.

Day, M. (1999). The scholarly journal in transition and the Pub Med Central proposal. *Ariadne* 21. [Online]. Available: http://www.ariadne.ac.uk/issue21/pubmed/intro.html [1 July 2004].

Day, M. (2001). Eprint services and long-term access to the record of scholarly and scientific Research. *Ariadne* 28. [Online]. Available: http://www.ariadne.ac.uk/issue28/metadata/ [1 July 2004].

Delamothe, T. *et. al.* (1999). Netprints: the next phase in the evolution of biomedical publishing. *British Medical Journal* 319: 1515-6. [Online]. Available: http://www.bmj.com/cgi/content/full/319/7224/1515 [1 July 2004].

eprints.org. [Online]. Available: http://www.eprints.org/ [1 July 2004].

Fecko, M.B. (1997). History and evolution of electronic resources. In Fecko, M. B. *Electronic Resources: access and issues* (1 – 14). East Grinstead: Bowker-Saur.

Fox, E. (1999). *Contribution by Edward A. Fox regarding Networked Digital Library of Theses and Dissertations (NDLTD)*. Paris: UNESCO Meeting, Paris, 27 – 28 Sep. [Online]. Available: http://www.unesco.org/webworld/etd/contributions/fox.rtf [4 June 2002].

Garner, J., Horwood, L. and Sullivan, S. (2001). The place of eprints in scholarly information delivery. *Online Information Review* 25 (4): 250 - 256.

Ginsparg, P. (1996). *Winners and losers in the global research village*. Invited Contribution, UNESCO, Paris, 19-23 Feb. 1996. [Online]. Available: http://xxx.lanl.gov/blurb/pg96unesco.html [1 July 2004].

Gonçalves, *et. al*. (2001). Flexible interoperability in a federated digital library of theses and dissertations. In *Proceedings of the 20th World Conference on Open Learning and Distance Education, "The Future of Learning – Learning for the Future: Shaping the Transition"*, *ICDE2001*, Düsseldorf, Germany, 01-05 April 2001. [Online]. Available: http://www.nudl.org/papers/ICDE2001.pdf [1 July 2004].

Gonçalves, M.A. and Fox, E.A. (2001).Technology and research in a global Networked University Digital Library. *Ciência da Informação* 30 (3): 13 - 23. [Online]. Available: http://www.ibict.br/cionline/300301/3030301.pdf [1 July 2004].

Goodrum, A. *et. al*. (2001). Scholarly publishing in the Internet age: a citation analysis of computer science literature. *Information Processing & Management* 37: 661 - 675.

Guedon, J. (2001). *In Oldenburg's Long Shadow: Librarians, research scientists, and the control of scientific publishing*. Annapolis Junction, MD: ARL. [Online]. Available: http://www.arl.org/arl/proceedings/138/guedon.html [1 July 2004].

Harnad, S. (1990). Scholarly skywriting and the prepublication continuum of scientific inquiry. *Psychological Science* 1:324-343. [Online]. Available: http://cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad90.skywriting.html [1 July 2004].

Harnad, S. (1998). Learned inquiry and the Net: the role of peer review, peer commentary and copyright. *Learned Publishing* 11 (4): 183-192. [Online]. Available: http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad98.toronto.learnedpub.html [1 July 2004].

Harnad, S. (1999) Free at last: the future of peer reviewed journal. *D-Lib Magazine* 5 (12). [Online]. Available: http://www.dlib.org/dlib/december99/12harnad.html [1 July 2004].

Harnad, S. (2000). The invisible hand of peer review. *Exploit Interative* 5. [Online]. Available: http://www.exploit-lib.org/issue5/peer-review/ [1 July 2004].

Harnad, S. (2001a). Research access, impact, and assessment. *Times Higher Education Supplement*, 1487:16. [Online]. Available: http://www.cogsci.soton.ac.uk/~harnad/Tp/thes1.html [1 July 2004].

Harnad, S. (2001b, April 26). The self-archiving initiative. *Nature*, 410: 1024 - 5. [Online]. Available: http://www.cogsci.soton.ac.uk/~harnad/Tp/nature4.htm [1 July 2004].

Harnad, S. and Hemus, M. (1998). All or none: no stable hybrid or half-way solutions for launching the learned periodical literature into the post-Gutenberg galaxy. In: Ian Butterworth (ed.). *The impact of electronic publishing on the academic community*. London: Portland Press. [Online]. Available: http://tiepac.portlandpress.co.uk/books/online/tiepac/session1/ch5.htm [4 June 2002].

Holtkamp, I. and Berg, D.A. (2001). *The impact of Paul Ginsparg's ePrint arXiv (formerly known as xxx.lanl.gov) at Los Alamos National Laboratory on Scholarly Communication and Publishing: a selected bibliography*. [Online]. Available: http://lib-www.lanl.gov/libinfo/preprintsbib.htm [1 July 2004].

Kling, R. and McKim, G. (1999). Scholarly communication and the continuum of electronic publishing. *Journal of American Society of Information Science* 50 (10): 890 - 906.

Kling, R. and McKim, G. (2000). Not just a matter of time: field differences in the shaping of electronic media in supporting scientific communication. *Journal of the American Society for Information Science* 51: 1306 - 1320. [Online]. Available: http://xxx.lanl.gov/ftp/cs/papers/9909/9909008.pdf [1 July 2004].

Kling, R., Spector, L. and McKim, G. (2002). *Locally controlled scholarly publishing via the Internet: the Guild Model*. Indiana: CSI, WP 02 - 01.[Online]. Available: http://www.slis.indiana.edu/csi/WP/WP02-01B.html [1 July 2004].

Krichel, T. (2000). Working towards an Open Library for economics: The RePEc project. *In* PEAK. *The Economics and Usage of Digital Library Collections*, Ann Arbor, Mi, 23 - 4, Mar. 2000. [Online]. Available: http://openlib.org/home/krichel/myers.html [1 July 2004].

Large, A., Tedd, L. and Hartley, R.J. (1999). *Information seeking in the online age: principles and practice*. London: Bowker-Saur.

Lawrence, S. (2001a). Free online availability substantially increases a paper's impact. In *Nature web debates*, 2001. [Online]. Available: http://www.nature.com/nature/debates/e-access/Articles/lawrence.html [1 July 2004].

Lawrence, S. (2001b). Online or Invisible? *Nature* 411 (6837), 521. [Online]. Available: http://www.neci.nec.com/~lawrence/papers/online-nature01 [1 July 2004].

Letnes, L. and Kelly, J. (2002). *AgEcon Search*: Partners build a Web resource. *Issues in Science and Technology Librarianship* 34, Spring 2002. [Online]. Available: http://www.istl.org/02-spring/article3.html [1 July 2004].

Luce, R. (2001). E-prints intersect the Digital Library: inside the Los Alamos arXiv. *Issues in Science and Technology Librarianship* 29, Winter 2001. [Online]. Available: http://www.library.ucsb.edu/istl/01-winter/article3.html [1 July 2004].

Luzi, D. (2000). Trends and evolution in the development of grey literature. The *International Journal on Grey Literature* 1 (3): 106 - 116.

McConnell, J. and Horton, R. (s/d) *The dawn of a New Era: The Lancet Electronic Research Archive in international health and e-print server*. [Online]. Available: http://www.thelancet.com/era [1 July 2004].

McKiernan, G. (2000). arXiv.org: the Los Alamos National Laboratory e-print server. *International Journal of Grey Literature* 1 (3): 127 - 138.

Meadows, A. J. (1998). *Communicating research*. San Diego, CA: Academic Press.

Moxley, J. (2001). Universities should require electronic theses and dissertations. *Educause Quarterly* 3: 61 - 63.

Moxley, J. *et. al*. (2001/2). *ETD Guide*, USF. [Online]. Available: http://etdguide.org/ [1 July 2004].

NATURE. *Future e-access to the primary literature web debate* 2001 - 2. [Online]. Available: http://www.nature.com/nature/debates/e-access [1 July 2004].

Needham, P. (2001). *The MAGiC Project – Interim Progress Report: October 2000 – August*. Cranfield: Cranfield University. [Online]. Available: http://www.bl.uk/concord/pdf_files/magicprogress.pdf [1 July 2004].

Nelson, M. (1999). *A Digital Library for the Advisory Committee for Aeronautics*. Hampton, VA: NASA Langley Research Centre. [Online]. Available: http://techreports.larc.nasa.gov/ltrs/PDF/1999/tm/NASA-99-tm209127.pdf [1 July 2004].

Nottingham ePrints. [Online]. Available: http://www-db.library.nottingham.ac.uk/eprints/ Nottingham ePrints. *About Nottingham ePrints*, Value added services.[Online]. Available: http://www-db.library.nottingham.ac.uk/ep1/information.html [1 July 2004].

OAI. *Open Archives Initiative*. [Online]. Available: http://www.openarchives.org/ OAI. Frequently Asked Questions. *What is the mission of the Open Archives Initiative.* [Online]. Available: http://www.openarchives.org/documents/FAQ.html [1 July 2004].

Odlyzko, A.M. (1995). Tragic loss or good riddance? The impending demise of traditional scholarly journals. *International Journal of Human-Computer Studies* 42: 71 - 122. [Online]. Available: http://www.dtc.umn.edu/~odlyzko/doc/tragic.loss.txt [1 July 2004].

Odlyzko, A.M. (2002). The rapid evolution of scholarly communication. *Learned Publishing* 15 (1): 7 - 19. Also to appear in *Bits and Bucks: Economics and Usage of Digital Collections*, W. Lougee and J. MacKie-Mason (eds.), MIT Press, 2002. [Online]. Available: http://www.dtc.umn.edu/~odlyzko/doc/rapid.evolution.pdf [1 July 2004].

Okerson, A. (1992). Publishing through the network: the 1990 debutante. *Scholarly Publishing* 23 (3): 170 - 177, quoted in Day, M. (1999). The scholarly journal in transition and the Pub Med Central proposal. *Ariadne* 21. [Online]. Available: http://www.ariadne.ac.uk/issue21/pubmed/intro.html [1 July 2004].

Okerson, A. and O'Donnell, J. (eds.) (1995). *Scholarly journals at the crossroads: a subversive proposal for electronic publishing*. Washington, D.C.: Association of Research Libraries. [Online]. Available: http://www.arl.org/scomm/subversive/index.html [1 July 2004].

OpCit: The Open Citation Project. [Online]. Available: http://opcit.eprints.org/ [1 July 2004].

Oppenheim, C. (1997). Towards the Electronic Library. In Scammell, Alison (ed.). *Handbook of special librarianship and information work* (397-407). 7th ed, London: Aslib.

Oppenheim, C. (2000). The Future of Scholarly Journal Publishing. *Journal of Documentation* 56 (4): 361 - 398.

OSI. *Budapest Open Access Initiative*. [Online]. Available: http://www.soros.org/openaccess/view.cfm [1 July 2004].

Pinfield, S. (2001a). How do physicists use an e-print archive? Implications for Institutional E-Print Services. *D-Lib Magazine* 7 (12). [Online]. Available: http://www.dlib.org/dlib/december01/pinfield/12pinfield.html [1 July 2004].

Pinfield, S. (2001b). Managing electronic library services: current issues in UK higher education institutions. *Ariadne* 29. [Online]. Available: http://www.ariadne.ac.uk/issue29/pinfield/intro.html [1 July 2004].

Pinfield, S. Gardner, M. and MacColl, J. (2002). Setting up an institutional e-print archive. *Ariadne* 31. [Online]. Available: http://www.ariadne.ac.uk/issue31/eprint-archives/intro.html [1 July 2004].

Prinsen, J. (2001). A challenging future awaits libraries able to change: highlights of the international Summer School on Digital Library. *D-Lib Magazine* 7 (11). [Online]. Available: http://www.dlib.org/dlib/november01/prinsen/11prinsen.html [1 July 2004].

Rowland, F. (1997). Print journals: fit for future? *Ariadne* 7. [Online]. Available: http://www.ariadne.ac.uk/issue7/fytton/intro.html [1 July 2004].

The SPARC Open Access Newsletter. [Online]. Available: http://www.earlham.edu/~peters/fos/index.htm [1 July 2004].

Schauder, D. (1994). Electronic publishing of professional articles: Attitudes of academics and implications for the scholarly communication industry. *Journal of the American Society for Information Science* 45 (2): 73 - 100.

Severins, T. *et. al*. (2000). PhysDoc – A distributed network of Physics Institutions Documents. *D-Lib Magazine* 6(12). [Online]. Available: http://www.dlib.org/dlib/december00/severiens/12severiens.html [1 July 2004].

Sompel, H. and Lagoze, C. (2000). The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine* 6 (2). [Online]. Available: http://www.dlib.org/dlib/february00/vandesompeloai/02vandesompel-oai.html [1 July 2004].

Suleman, H. *et. al*. (2001a). Networked Digital Library of Thesis and Dissertations – Bridging the gaps for global access – Part 2: Services and research. *D-Lib Magazine* 7 (9). [Online]. Available: http://www.dlib.org/dlib/september01/suleman/09suleman-pt2.html [1 July 2004].

Suleman, H. *et. al*. (2001b). Networked Digital Library of Thesis and Dissertations – Bridging the gaps for global access – Part 1: Mission and Progress. *D-Lib Magazine* 7 (9). [Online]. Available: http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html [1 July 2004].

Tenopir, C. and King, D.W. (2000). *Towards electronic journals: Realities for scientists and publishers*. Washington, D.C.: Special Libraries Association.

Walker, T. (1998). Free Internet access to traditional journals, *American Scientist* 86 (5). [Online]. Available: http://www.sigmaxi.org/amsci/articles/98articles/walker.html [1 July 2004].

Walker, T.J. (2002). Two societies show how to profit by providing free access. *Learned Publishing* (forthcoming). [Online]. Available: http://csssrvr.entnem.ufl.edu/~walker/epub/ALPSPmsDS2.pdf [1 July 2004].

Warr, W. (2001). *A report on the presentation "Chemistry Preprint Server: a revolution in Chemistry Communication*, Spring 2001 National ACS Meeting (CINF Division) San Diego, USA. [Online]. Available: http://www.chemweb.com/docs/cps/cps.pdf [1 July 2004].

Wilkinson, S. (2000). Chemistry Preprints Come Into Vogue. *Chemical Engineering News* 78(23): 15-6. [Online]. Available: http://pubs.acs.org/isubscribe/journals/cen/78/i23/html/7823notw5.html [4 June 2002].

Information Discovery and Retrieval Tools

Michael T. Frame
U.S. Geological Survey, Center for Biological Informatics
Mail Stop No. 302
12201 Sunrise Valley Drive
Reston, Virginia 22092
USA

mike_frame@usgs.gov

ABSTRACT

Due to the rapid growth of electronically accessible content from the Internet, there is a corresponding increase in demand for information of all types from a number of diverse users. Although the World-Wide Web presents tremendous opportunities to users for access to this wealth of information, the quantity of that information can be overwhelming. The user who attempts to find information can become confounded by the sheer volume of data and information returned as "pertinent" to his/her need. In addition, current awareness becomes an obstacle, as variations in search engine crawls of the Web, as well as the user's own ability to keep up with frequent queries to multiple search tools, can prevent timely access to and knowledge of pertinent information. This session will focus on the various Internet search engines, directories, and how to improve the user experience through the use of such techniques as metadata, meta-search engines, subject specific search tools, and other developing technologies.

1.0 BACKGROUND

Ever since the Internet's beginnings in the 1990s, the amount of information available on the World-Wide Web has steadily increased. It is estimated that close to 10 billion web pages exist on the World-Wide Web today. As expected, this number is continuing to grow; however, at a much slower and some say more controlled rate. The rate of growth of World-Wide Web content has also caused the community of casual and advanced users, to consider alternative means to finding information.

As the information content has grown on the World-Wide Web, so too has the need for improved tools and products to aid users in this discovery of information. Several tools basically perform the same function, but may differ slightly in their methods and results. This primarily has to do with vendor specific interpretation of World-Wide Web terms such as: Spam, spider/crawler configurations, and collection size. All of this leads to industry estimates that less than 20% of the entire content of the World-Wide Web is available to the typical user (World-Wide Web Consortium 2004). This paper investigates various terminologies and provides simple techniques users can perform to improve their search experiences on the World-Wide Web.

2.0 BASIC TERMINOLOGY

2.1 What Do Internet Search Engines Really See?

From a user's perspective, as shown in Figure 1, users often simply enter a term in a simple search box and wait for results. They are oblivious to what the computer or system is doing. This is the way it should be. If users have to worry about how an Internet search engine is configured or what it expects, then most likely

the search engine user interface needs to be redesigned or another product selected. Users have too many other things to do, whether at work or home, to concern themselves with learning the various idiosyncrasies of each Internet search engine.

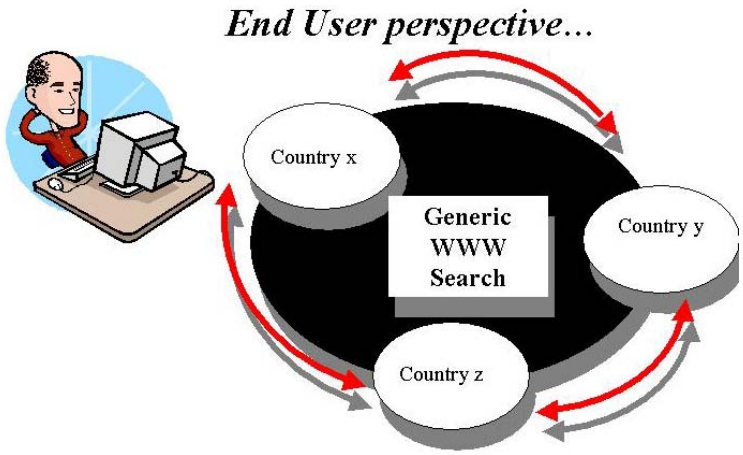


Figure1: Typical User Search.

However, what the user often does not realize is that Internet search engines primarily read the underlying document codes or "metatags" within a document. Metatags are document tags or properties that are often stored within the Header of an HTML document or within the document itself. Figure 2 below describes a typical view that an Internet search engine would see when it indexes a document.

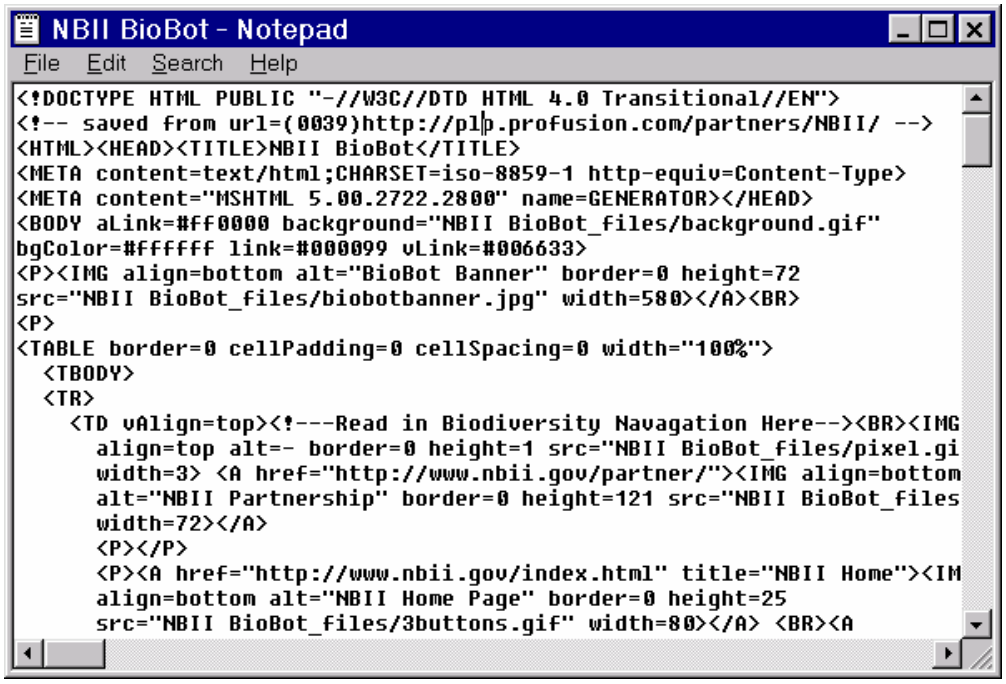


Figure 2: Typical Internet Document as Viewed by Search Engines.

2.2 What is Spam?

"Spam" is a term you often hear thrown about on the World-Wide Web today. Spam is not just a popular Hawaiian luncheon meat anymore. Understanding what spam is and is not is very important in understanding how search engines on the WWW discover and display information to users. Spam is considered to be anything that a software developer or HTML creator does to try to falsify his or her content to a web engine. In today's web environment content creators jockey for position on Internet search engines results/hits lists and often resort to categorizing their sites in ways that may not truly represent the content or overall purpose. This is considered spamming a search engine crawler or data harvester. Tricks commonly employed by web content creators include applying keywords within the Header section of an HTML document that have nothing to do with their site, or simply creating BLANK HTML pages with white text so that users don't see the content, but a search engine can. Internet Search Engines are all wise to these tricks and this is why it is often difficult for content producers and/or developers who have truthful content and are trying to do a good job in making their content available understand what an Internet search engine expects and applies preferences to.

2.3 The Basic Internet Search Engine Model

Internet search engines on the WWW "harvest" data from publicly available web sites via automated jobs or crawls. This harvesting or gathering of summary information (usually items such as URL, keywords, summary description) to a central point is done with spiders and/or crawlers. Spiders and crawlers are simply automated jobs or processes that run from an Internet search engine provider's server and scour the WWW for content. This content is then made available through the Internet search engine providers' central index. Figure 3 below demonstrates this process.

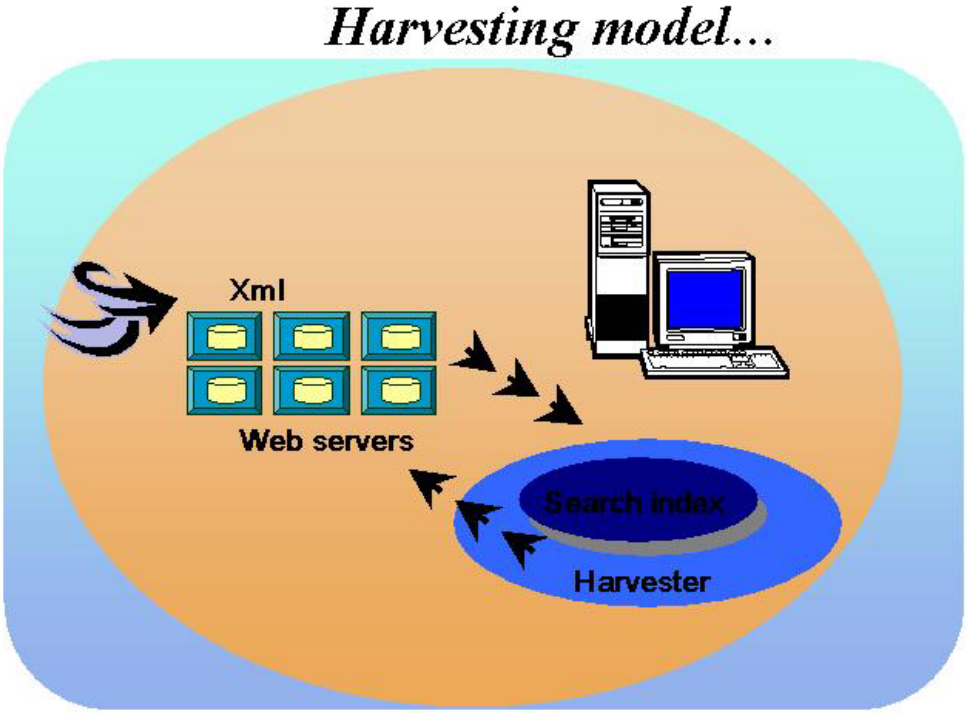


Figure 3: Basic Internet Search Engine Harvesting Model.

2.4 What are Metatags and Why are they Important?

Embedding metatags within the HTML of your Web site not only promotes higher rankings, and thus, better retrieval, of your site by many of the major search engines, but also provides a foundation for future information retrieval and discovery on the Web as the web evolves into a more structured organization of content. The algorithms used by search engines constantly change; however, the presence of metatags on your pages can often make a dramatic difference in enabling users to find your information. Remember, too, that as various sites apply metatags, an **integrated** system whereby users can easily locate your site through a search engine are likely to explore other related sites within the WWW.

The table below describes both standard metatags and unique discipline, in this example biological information, metatags that all can be implemented on web sites. Search engines require some tags, while others are optional, depending upon the scope and context of the page(s) under development. Additional metatag requirements may be added as retrieval tools become more sophisticated. Fortunately, the creation and editing of metatags is a quick and simple process, thanks to the development of metatag software, which can rapidly generate tags selected by a content provider across designated pages, directories, or an entire site.

The metatags in Table 1 below are all standard HTML 3.0 or above supported tags. If users are using dynamically created web sites, the metatags described below can simply be created automatically out of a database dump or export.

Table 1: Recommended Metatags

Metatag	Definition	Format & Sample Value
Author	The Author Tag contains name of the content provider (*not* the Webmaster / programmer).	<meta name="author" content="Bob Johnson">
Title	Even though the Title tag is not considered a true metatag, it is critical in search engines' ranking algorithms, and provides users with general information about your page. Search Engines results/hit lists also display the Title tag. Up to 80 characters can be contained within this tag.	<TITLE> West Nile Virus: Wildlife Impacts - NBII</TITLE> *** please maintain this format when naming your pages ***
Keywords	Keywords are probably the most important meta-tag that a Web site manager can include. Up to 1000 characters can be contained within this tag. *Your keyword contents should include the basic tags at left, plus all terms relevant to your site and particular sub-sections. Include several generic terms that apply to your entire node, plus terms specific to various sub-directories and pages. Try to think of as many synonyms for your terms as you can. Note that you need to include term variations (e.g. bird, birds, birding, birdwatcher), as the search engines do not employ stemming when parsing keywords. Spelling counts! Use terms found within the page contents to boost relevancy rankings.*	<meta name="keywords" content="*your page-specific keywords….,* NBII, National Biological Information Infrastructure, biology, biodiversity, natural resources, reference, education, ….. ">> *place these standard keywords AFTER your page-specific keywords*
Page Description	The Description tag is used by search engines to display information about your page and to index its contents. Up to 200 characters can be contained within this tag. The description often determines whether the searcher will choose to view your page. Make the description relevant to the particular sub-section or page; don't rely on one generic description for all pages on your site. Use keyword tag terms in your description to boost term relevancy rankings.	<meta name="description" content="This is the textual description for your page. Please make sure your spelling is correct and include any relevant keywords within the Description tag.">
Language	Even though most content on the web is in English, the Language tag adds value to your Web site, helping users limit search engine retrieval to a particular language.	<meta name="language" content="en-us">
Classification	The Classification tag is often used by a number of the Web search engines when you register your site and/or when your site is indexed so that your site can be classified with other similar sites. Typical values include: "Government, Science, Education, etc."	<meta name="classification" content="Government, Science">
Ratings/PICS	The Ratings and PICS tags are used by Internet providers and search engines to limit access to a particular page. Often this is used to restrict access to "Mature Audience Only" pages for children using the Internet. Typical Values include: "General, Restricted, Mature, Safe for Kids", etc. Because filters are becoming more common within retrieval tools and browsers, or as added software, these tools may arbitrarily block your site if the tag is not implemented.	<meta name="rating" content="General, Safe for Kids">

Table 2 below describes the unique or custom metatags for a domain specific organization. In this case, these custom metatags are relevant to categorizing, displaying, and delivering biological data and information.

Table 2: Domain Specific Metatags (Custom Tags)

Metatag	Definition	Format & Sample Value
Species Scientific Name	The Scientific Name of a particular Species on the web page being classified. NBII Partners are strongly encouraged to utilize the Integrated Taxonomic Information System (ITIS) (http://www.itis.usda.gov/plantproj/itis/index.html) as its basis for completing this information.	<meta name="Species Scientific Name" content="Parnassius smintheus">
Species Common Name	The Common Name of a particular Species on the Web page being classified. The Common Name is extremely important to both expert and novice users for finding information about a particular species. ITIS is a source for completing this meta-tag.	<meta name="Species Common Name" content="Rocky Mountain Parnassian">
Organization	The lead Partner organization that maintains the specific Web site/page being classified. The use of standard controlled lists is strongly encouraged for completing this field.	<meta name="Organization" content="USGS Center for Biological Informatics">
Web-site Theme	The high-level Theme (Education, etc.) that your Web page falls under within a web structure.	<meta name="website Theme" content="Education">
Web–site Category	The specific Category, within the website Theme, that your Web page falls under.	<meta name="website Category" content="General Curriculum">

Domain specific metatags greatly aid a particular community of users in the discovery and identification of quality resources. For example, if a user accesses one of the search engines on the World-Wide Web today and searches for a specific bird, i.e. "common loon", the search result produces a hit list of more than 13 million results. Some of these results are most likely pertinent to the user, but most are not and it is infeasible for a user to navigate through 13 million web pages for relevant data.

To resolve this issue, programs such as the National Biological Information Infrastructure (http://www.nbii.gov) have been implementing a refined and improved spidering methodology with its partners and applying metatags within its local and partner pages. As a result, users can now easily narrow their results lists to 62,000 web pages with the same search that yielded over 13 million results. These spidered and indexed pages are primarily biological in nature and due to the intellectual effort that is currently ongoing within the NBII Program for adding information content to the NBII System, users can expect to receive more targeted and a higher quality result than directly access the WWW and its search engines. Users also have the ability to narrow their search results to 1,400 web pages and information sources through the direct querying of meta-information contained within a domain specific or custom meta-tag called "Common Name". As one can imagine, this saves users tremendous time and presents authoritative and related information to a user without requiring an already information overloaded user to review a large number of primarily non-pertinent results.

3.0 TYPICAL SEARCH ENGINE FEATURES AND CAPABILITIES

As stated, all search engines are mostly the same, but often different in their implementation and configurations. Below are some of the features you would expect to find in a typical search engine. Often low-end search engines may or may not have all of the features noted or may be limited in how many documents one may index or limited on the size of your collection.

- Contains an automated spider or crawler
- No theoretical limits in the amount of indexing (limited by hardware)
- Supports remote indexing
- Continual background indexing of content
- Custom metatag support (some low-end products do not support this feature)
- Support for indexing PDF, .doc, etc. (some low-end products do not support this feature)
- Supports URL and word exclusions & inclusions
- SSI supported
- Search by custom metatags
- Case sensitive or insensitive searching
- Simple search interface
- Ability to customize search results pages
- Boolean Searching capabilities
- Provide users meta description and page title in search results
- Inexpensive cost, – $200
- Easily customizable search/results interface
- Result weighting feature
- URL Inclusion list for target indexing
- Require significant memory (RAM) and disk space as the collection grows
- Low-end alternatives often do not possess the capabilities to do phrase or natural language searching.

4.0 WHAT CAN YOU DO AS A CONTENT DEVELOPER OR SOFTWARE DEVELOPER TO IMPROVE DISCOVERY OF YOUR CONTENT?

Users can do several things to help ensure that their information content is more readily found on the WWW today. Some of these things make perfect sense, but users often do not dedicate the necessary resources required to make them happen on a regular basis. Each environment and web site is different; however, the general principles and techniques noted below will help any web content producer.

- Implement metatags on your and your partners web sites
- Update content frequently
- Register your site with the major search engines (tools exist to aid in this process)
- Perform a basic study of where your site results within the major search engine providers

- Do not spam the search engine providers
- Re-evaluate your web site directory structure to ensure information is appropriately categorized/ described within your URL strings
- Look through your server log files to determine what users are trying to find on your site and/or the path they are using to find information
- Perform basic usability testing of your site to determine what users expect and can easily gather from your site. This also may determine why users go to an Internet search engine provider versus accessing your site directly.
- Realize that Internet search engines don't all act the same, index at the same time period, and often value a particular metatag, document date, etc. more than another vendor product.

5.0 CONCLUSION

As one can see, maintaining awareness and improving delivery of your information via the WWW in today's environment is almost a full-time job. As Internet search engine providers become more sophisticated, so too will it be necessary for content producers and providers to restructure their information to take advantage of such capabilities. With the advent of new technologies, such as XML and SOAP, information content will be more readily able to be delivered at a more granular scale and to a more targeted audience. However, these technologies are still in their infancy, as it comes to the overall web content, and Internet search engines will continue to be one of the major sources whereby users access to gather information.

6.0 REFERENCES AND FURTHER READING

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., and Ragahavan, S. (2001). Searching the web. *ACM Transactions on Internet Technology* 1(1): 2-43.

Kobayashi, M. and Takeda, K. (2000). Information retrieval on the web. *ACM Computing Surveys* 32(2): 144-73.

Nobles, R. and O'Neill, S. (2000). *Streetwise maximizing web site traffic: build web site traffic fast and free by optimizing search engine placement.* Avon, MA: Adams Media Corporation.

Notess, G. (2004). *Search Engine Showdown Homepage*. [Online]. Available: http://www.searchengineshowdown.com [5 July 2004].

President's Committee of Advisers on Science and Technology (PCAST) Panel on Biodiversity and Ecosystems. (1998). *Teaming with life: investing in science to understand and use America's living resources.* [Online]. Available: http://www.nbii.gov/about/pubs/twl.pdf [9 May 2002].

Sullivan, D. (2004). *Search Engine Watch Homepage*. [Online]. Available: http://searchenginewatch.com [5 July 2004].

World-Wide Web Consortium Homepage. (2004). [Online]. Available: http://www.w3c.org [5 July 2004].

Yonaitis, R. (2000). *The elements of <web site> promotion.* Concord, NH: Hiawatha Island Software Corporation.

Electronic Collection Management and Electronic Information Services

Gladys Cotter
U.S. Geological Survey
12201 Sunrise Valley Drive
MS 300
Reston
Virginia 22092
USA

gladys_cotter@usgs.gov

Bonnie Carroll, Gail Hodge *and* Andrea Japzon
Information International Associates, Inc.
1009 Commerce Park Dr.
Suite 150 / P.O. Box 4219
Oak Ridge, TN 37830
USA

bcarroll@infointl.com, gailhodge@aol.com,
ajapzon@pop200.gsfc.nasa.gov

ABSTRACT

As the life cycle of information products has become increasingly digital from "cradle to grave," the nature of electronic information management has dramatically changed. These changes have brought new strategies and methods as well as new issues and challenges. At the bottom line the services are increasingly delivered to a desktop from distributed publishers or information providers. Information organizations act either as primary information providers or as brokers between the user and the primary service provider. There has also been a significant reorientation from "ownership" of materials to "access" to information. This paper covers developments in the factors and strategies affecting collection management and access. It discusses major trends in electronic user services including electronic information delivery, information discovery and electronic reference. Finally, it addresses the challenges in user and personnel education in response to this electronic environment and an increasingly information literate user population.

1.0 INTRODUCTION

One of the earliest significant implementations of electronic information management for libraries and information centers began in the 1960s and 1970s with dialup access to remote electronic databases such as those provided by Dialog and LexisNexis. The Internet has further contributed to the success of remote information systems and databases by increasing the information transfer rates from 300 baud to multiple megabits. The World Wide Web simplifies the process and interpretation of the bits that are transferred.

With these increased technical capabilities, new online databases that provide bibliographic and full-text access to information resources have proliferated and the volume of electronic information content now available from the desktop is staggering (Lyman and Varian 2000) – and it was recognized twenty years ago that the volume of information was already beyond our ability to absorb the increase (de Sola Pool 1983). Library catalogs have been computerized[1] and are available and interoperable across the Internet. Audio, video, and multimedia resources are available, as are interactive services from games to banking. Filtering and push/pull delivery services help us manage information proliferation.

[1] Michael Buckland (1997, ch. 5) makes the very useful distinction between electronic libraries and automated libraries. Electronic libraries contain electronically stored documents. Automated libraries (which may or may not be also electronic libraries) use automated search and retrieval systems, for example, online public access catalogs (OPACs).

Electronic collection management and electronic information services are being pushed by the dramatic increase in the amount of digital information from an increasing variety of sources, by the new technologies in the information field, and by heightened user expectations. This paper explores some of the theoretical and practical aspects of collection management in the digital age. It also looks at the major trends in electronic user services including electronic information delivery and electronic reference. Finally, we address the challenges in user and personnel education in response to this changing environment and the increasingly information literate user population.

1.1 The Digital Revolution in Libraries and Information Centers

As a key institutional structure for providing information "collections" and "services" to users, the library (throughout this paper, "library" represents libraries, information centers, and special collection builders) has had to respond rapidly to a changing external publishing environment. The Internet and most recently the World Wide Web have impacted libraries and their delivery of service in ways unthought of except by visionaries like Vannever Bush (1945) or H.G. Wells (1937). Libraries continue to fulfill their mandates but in ways different than they once did. "Library" is less a place than a concept that represents certain processes or services (Birdsall 1994). If we accept that "library" is something other than place, the continuity of library function is maintained even as libraries transition from, in Negroponte's (1995) terms, atoms to bits.

We now distinguish between three different types of libraries: traditional libraries, digital (electronic or virtual) libraries, and hybrid libraries. Traditional libraries have physical objects.

Digital libraries are byte-based. According to the Digital Library Federation, digital libraries have very much the same form and function as traditional libraries:

> Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities. (Digital Library Foundation http://www.clir.org/pubs/issues/issues04.html#dlf).

Michael Lesk (1997) suggests that digital libraries share three common traits: (1) they can all be searched, (2) they can all be accessed from anywhere, and (3) they can all be copied using electronic means without error. Are then digital libraries traditional libraries but more so?

A number of libraries have emerged that are born digital. Digital libraries, including Web-based collections, are redefining both the role of electronic information storage and retrieval as well as the role of traditional libraries. Digital libraries include collections of books, journal articles, graphics, newspapers, and other material in digital format; in sum, collections of digitized content.

As a practical matter, most libraries must manage both traditional "physical objects" and digital materials in a hybrid library environment (Pinfield et al 1998). In certain disciplines, such as astronomy and physics, legacy information that may never be digitized can be as important as the information that is only born digital. Many libraries have begun the transformation from traditional to hybrid forms (for an example, see Bjoernshauge 1999). Many academic and public libraries have incorporated digital material into their collections if only by adding links to that material in their OPACs or from library web sites. Depending on who is speaking, hybrid libraries may be the model for the foreseeable future (Leggate 1998) or a way-station to another form (Oppenheim and Smithson 1999). The HyLiFe Hybrid Library Toolkit (2002) provides guidance to those seeking to develop or migrate to hybrid libraries.

What is the future evolution for electronic libraries? Harold Billings argues that research libraries are becoming huge linked relational libraries. He sees the need for formal relationships among existing libraries to build the mega-libraries that would meet anticipated user needs. Libraries might simply evolve into gigantic and, hopefully, well-organized portals or, as Lancaster and Warner (2001) describe, one model as "switching stations." On the other hand, it is possible to see the impact of "customization" in "MyLibraries", which are desktop entities that allow a user, either individually or through profiling, to select his resources of interest and how they will be displayed. Some discipline- or domain-specific "MyLibraries" have been constructed. Most academic libraries provide some form of "MyLibrary" service to their patrons (Billings 2000).

Regardless of the path for future development, large portals or customized MyLibraries, the electronic library, if it is to have meaning, must bring value to collection management and services from that collection. Atkinson (1993) emphasizes the need for organization, selection, and management of resources in digital libraries, just as these resources are managed in traditional libraries. Chen states that for digital materials to be a library, they must be organized according to some standard (1998). Library and commercial portals provide user-centered and enhanced access to information resources by evaluating and selecting local and global information that is context driven. As with print resources, librarians designing portals are careful to provide access to only the most authoritative electronic sources to create trustworthy access to information. (Mischo 2001) The National Science Digital Library (NSDL) and Science.gov are examples of two science portals created to provide access to science information that is selective and authoritative in a way that search engines such as Google can never be (Tennant 2003a). To do so, there must be an articulated collection management policy based on sound philosophical underpinnings.

2.0 ELECTRONIC COLLECTION MANAGEMENT

The electronic library and collection management in that library are relatively new concepts. The earliest literature dates to the 1970s and 1980s (Harter and Kister 1981, Dowlin 1984). In 1984 Kenneth Dowlin (1984, p. 33) suggested that the electronic library has four attributes: (1) "management of resources with a computer," (2) "the ability to link the information provider with the information seeker via electronic channels," (3) "the ability for staff to intervene in the electronic transaction when requested by the information seeker," and (4) "the ability to store, organize, and transmit information to the information seeker via electronic channels." An electronic library utilizes both electronic information resources and electronic means to manage and move those resources.

There are sound reasons to consider collection management in an electronic environment. Libraries bring more than organization and intermediation to information collections. They also bring authority. Inclusion in a collection implies pertinence and appropriateness. At the same time, the same information "content" can and will be provided in different "containers."

2.1 The Key Challenge: Ownership versus Access

The move to electronic information management has resulted in a number of debates. "Ownership versus access" has been one of the more important issues. Budd and Harloe (1994) distinguish between the ownership-based and the access-based organization. In the former, emphasis is placed on building "on-the-shelf" collections while, in the latter, emphasis is placed on access to resources, regardless of where they are "owned." Value is assessed differently. For the traditional model, the value of a collection is its size. For the access-based library, value is defined as the ability to retrieve useful information. The former library collects "just in case" material is needed; the latter provides it "just in time."

Both models contain pitfalls and problems. Keller (1992b), for example, has argued that "[n]ew access instead of ownership paradigm leads ultimately to an environment where 'all is meta information,' with no or few ideas on the shelves." The issue of access also brings in a whole new set of questions regarding archiving and preservation, intellectual property including fair use, as well as conditions for purchase which have moved to a complex set of licensing terms and conditions. Buckland (1997) suggests that libraries consider *ownership* for high demand items and *access* for those in low demand.

Given the spread of digital access, union catalogs, and universal borrowing, it is no longer so important what an information organization contains (owns); rather, the focus is on the services (access) the organization can provide (Ferguson and Kehoe 1993). This change in a basic tenet of library management has resulted in the need for different library use metrics.

2.2 Access Models

"Access" to digital information comes through several modes of access. Each of these forms can be considered as part of the "ownership versus access debate" and impacts the new ways of managing electronic collections. The following describes four models now employed in the digital environment. These are the interlibrary loan model, the universal borrowing model, the fee-based model, and the no fee model. The interlibrary loan (ILL) model has been with us for many years. Universal borrowing (UB) is a recent phenomenon first seen in the mid-1990s. The fee-based electronic access model dates to the early 1960s with the advent of electronic database services like Dialog. Today, a large volume of "no fee" or "free" services are Web-based and emerged as major resources in the mid-1990s.

2.2.1 Interlibrary Loan Model

Interlibrary Loan (ILL) is a process by which one library borrows from other libraries materials it does not hold in order to meet the information needs of its patrons. Interlibrary Loan is not a new concept nor is it one that emerged out of the digital revolution. ILL is however facilitated by various online services including electronic union catalogs (like OCLC's *WorldCat*) and automatic ILL request services attached to OPACs and online databases. The ILL community has developed a continuing interest in using the Web and other means to facilitate the ILL process.

The North American Interlibrary Loan and Document Delivery (NAILDD) Project promotes the development of efficient ILL/DD delivery systems using networked technologies. NAILDD has identified three areas of primary concern: "comprehensive and flexible management software, improvements in ILL billing and payments, and system interoperability via use of standards" (Jackson 1998). OCLC has played a major role in developing system interoperability, facilitating billing and financial transfers (IFM or ILL Fee Management), and development of management software. A number of international initiatives led by the Research Libraries Group (RLG), the Library Corporation (TLC), Ameritech Library Services (ALS), AGCanada, and others have sought to improve system interoperability and information flows, thus enhancing digital access.

OCLC manages an international Interlibrary Loan Service or Global Sharing Group Access Capability (GAC), built upon its union catalog WorldCat. It utilizes a standard Web interface and software (http://www.oclc.org/services/brochures/).

Many countries have developed model codes for ILL, for example the American Library Association – Reference and User Services Association (ALA-RUSA) Interlibrary Loan Code for the United States (http://www.ala.org/rusa/stnd_lnc.html). ILL exchanges among countries are guided by the International Federation

of Library Associations and Institutions (IFLA) International Lending: Principles and Guidelines for Procedure (http://www.ifla.org/VII/s15/pubs/pguide.htm).

The Interlibrary Loan system is guided by a set of standards (ISO 10160 and ISO 10161). These standards were developed to insure interoperability among electronic ILL systems and their application protocols. These standards and protocols are managed by the ILL ISO Maintenance Agency. The National Library of Canada serves as host.

2.2.2 Universal Borrowing Models

The Universal borrowing models (UB) allow authorized users from one system to borrow (access collections) from libraries within a consortium. (The term "universal" in this context refers to providing access to everyone within a defined group, not universal in the sense of totally open.) There are two major models for UB arrangements. In the first, libraries of different types within a common jurisdiction permit intra-jurisdictional lending. This is used primarily within a particular geographic area, such as a county or state library network, that includes public, academic, and special libraries. The second type of UB involves libraries of the same type such as academic research libraries. Some large libraries may belong to multiple groups.

In the digital world, consortia or other pre-coordinated groups of organizations are increasingly active and pervasive due to the need to get the most favorable conditions under licensing agreements. Cost models for publishers of digital information are in serious flux and the need for groups that build collections to work together in their dealings with publishers and in developing access infrastructures for digital collections has become increasingly important.

2.2.3 Fee-Based Access

Although the Web initially encouraged free and freely available information, as it matures, commercial publishers are actively using it for vending their electronic material. The volume of fee-based access will continue to increase at an increasingly rapid rate.

Examples of the increasing number and variety of resources available as fee-based services include access to bibliographic and fulltext databases, to online journals, and to electronic books.

- There are a number of electronic databases that provide bibliographic and/or fulltext access to documents. These include services like Dialog, LexisNexis, Westlaw, Ovid, Chemical Abstracts' STN, OCLC's *ContentsFirst*, OCLC's *FirstSearch,* CARL UnCover, British Library Document Supply Centre Inside Information and Inside Conferences, ISI Current Contents, ISI – The Genuine Article, and the Canadian Institute for Scientific and Technical Information (CISTI).
- There are a number of database providers, like the Thomson Companies, that own a wide array of properties. Their holdings range from the *Physicians Desk Reference* and *Jane's Warships*, to the ISI collections. They also maintain copyright, patent, and trademark databases.
- There are a growing number of publishers providing direct individual or library subscriber access to online e-publications. E-publications either replace or supplement the paper version of the journal. There are a variety of models for the publication of e-publications, with some retaining the traditional periodicity of "issues" while others are providing almost continuous updates as articles become available.
- E-journals and hybrid-journals provide journal access either by subscription or association membership. These include *Science* and *Nature. The Journal of the American Society for Information*

Science and Technology is offered to members in either paper or electronic format (or both for an additional fee). Companies like MCB University Press and Elsevier offer bundles of online journals, multiple titles for a single subscription price, to libraries and individuals.

- E-book providers are Web based vendors of online and for the most part popular books. They are fee-based services charged to the consumer. There are a number of services that provide access to e-books. See, for an example, eBooks (http://www.ebooks.com/). Adobe provides pointers to e-book vendors (http://www.adobe.com/epaper/ebooks/ebookmall/main.html).

2.2.4 No Fee Electronic Access

No fee access to digital materials has become increasingly available through the Internet and, thereby, directly to the end user. Materials can be read online or downloaded in a variety of formats, including pdf, Microsoft reader, and html. From the "collection" point of view, providing identification and access to the free sites brings with it a number of issues. Since these are free materials and the level of responsibility of the "publishers" may vary considerably, collection developers have the difficult challenge of determining whether digital access provides sufficient continuity in their collection development scheme. Maintaining links to free electronic sites is a major collection maintenance challenge for digital collections. One major source of free material that is rather durable and generally of good provenance is US government material, especially policy and technical documents because US government materials cannot be copyrighted. Some sample sites for free material are:

- Project Gutenberg (http://promo.net/pg/), begun in 1971, permits its users to download books.
- Online Books (http://onlinebooks.library.upenn.edu/) at the University of Pennsylvania provides portal access to more than 15,000 e-books. It provides pointers to materials in its collection and on the servers of other providers. The text is marked up to provide cross references by hypertext links to material on the same subject in the collection.
- Subject gateways are comprehensive collections of digital and often Web documents organized around a set of central themes. Examples include the WWW Virtual Library (http://vlib.org/), BUBL LINK (http://bubl.ac.uk/link/), the Internet Guide to Engineering, Mathematics and Computing (http://www.eevl.ac.uk/) WebMD (http://www.webmd.com), and so on. Preserving Access to Digital Information or PADI (http://www.nla.gov.au/padi/) is a subject gateway to digital preservation issues. The Resource Discovery Network (http://www.rdn.ac.uk) is a metagateway, with links to major gateway sites.
- Governments provide Web based database access to a wide range of information. Examples include Thomas, a Library of Congress gateway to US Congressional documents (http://thomas.loc.gov/) and Edgar, a service of the Security and Exchange Commission (http://www.sec.gov/edgar/searchedgar/webusers.htm). Two of the major US science agencies, the U.S. Department of Energy and NASA provide free access to large technical report collections. See http://www.osti.gov/ for DOE and http://ntrs.nasa.gov/ for NASA.
- E-print and pre-print archives -- Los Alamos National Laboratory broke new bibliographic ground when it established an e-print archive (http://arxiv.org/ and mirrored at: http://xxx.lanl.gov/). This archive and others like it have proved invaluable in fields with fast breaking innovation.
- Many e-journals and some h-journals (hybrid, or journals published in paper and electronically) offer free access to their articles. These include a number of popular and scholarly journals offered in electronic format without charge or subscription; for example, the venerable *Scientific American* (http://www.sciam.com/), as well as the information science journals *Information Research*

(http://informationr.net/ir/), *D-Lib Magazine* (http://www.dlib.org/), *FirstMonday* (http://www.firstmonday.dk), and *Ariadne* (http://www.ariadne.ac.uk/). Many newspapers offer free access to all or parts of their editions. These include *El Día*, *New York Times*, *Wall Street Journal*, *Le Monde*, *Helsingin Sanomat*, *The Times of India,* to name a few.

A major movement on the part of scholars, public advocates, and some publishers would greatly extend the number of e-journals and other materials available under the free access model. The Open Access Movement asserts that scholarly materials, particularly those in the sciences, should be freely available to all. "By 'open access' to this literature, we mean its free availability on the public internet, permitting any user to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited. Open access eliminates two kinds of access barriers: (1) price barriers and (2) permission barriers associated with restrictive use of copyright, licensing terms, or DRM [digital rights management]." (Budapest Open Access Initiative, 2002 (http://www.soros.org/openaccess/read.shtml)).

The list of "open access" materials, particularly e-journals is increasing quickly. The Directory of Open Access Journals, maintained by Lund University Libraries and sponsored by the Information Program of the Open Society Institute and SPARC (Scholarly Publishing and Academic Resources Coalition), includes over 830 open access journals in 15 subject categories as of April 2004 (http://www.doaj.org). Some of these journals are alternatives to the more expensive commercial journals in various disciplines developed by open access publishers such as BioMed Central, the SPARC partners, and BioOne. These organizations may also act as trusted third parties for other publishers who are willing to deposit their materials in an open access arrangement with terms and conditions.

There have been several major statements of support for such a movement by scholars, practitioners, and even governments. These include the Budapest Open Access Initiative (http://www.soros.org/openaccess/read.shtml), the Bethesda Statement on Open Access Publishing (http://www.earlham.edu/~peters/fos/bethesda.htm) and the Washington DC Principles on Free Access for Science (www.dcprinciples.org). The latter statement seeks to achieve a middle ground, promoting free access while ensuring the sustainability of the scientific, technical and medical publishers. While there has been little movement in this direction from the physical sciences, "open access" is gaining momentum among the biomedical community. Key activities include the Public Library of Science, BioOne, BioMed Central and PubMedCentral. Peter Suber's web blog, *Open Access News*, is a comprehensive blog of announcements, news and articles on this topic (http://www.earlham.edu/~peters/fos/fosblog.html).

2.3 Selection of Information for Electronic Collections

The number, scope, and type of information resources in electronic and print format, as Lyman and Varian (2000) make clear, are overwhelming. There are many sources from which information can be drawn, and there continues to be a need to effectively evaluate those resources. Libraries have long functioned as one of the chief mechanisms for evaluation of information quality and relevancy.

If we take to heart Birdsall's (1994) conception of "library" as process or service rather than "place," libraries must put added focus on how to manage collections. Electronic collection development must be consistent with the mission and an overall collection development plan. At the same time, collection development plans

should take into account the electronic resources now available to libraries (Gessesse 2000). As it becomes easier or more efficient to move electronic and physical objects from a collection repository to the end user, the logic of maintaining redundant collections declines. Through selective collection policies, scarce resources can be used to broaden collections rather than to duplicate them.

In Lee's (2002) *Electronic Collection Development: A Practical Guide,* he discusses the application of print selection techniques to electronic media. Evaluation criteria specific to electronic tools are well-covered including interface usability, remote authentication, and archiving.

There are a number of online aids to assist librarians in collection development for either electronic or "traditional" materials. Librarians have discovered that the online booksellers like Amazon provide a vehicle for useful reviews and for purchasing.

With quality and permanence caveats in mind, the Web can be a source for collection development (see Lee 2000). Web material should be subjected to the same scrutiny that any other resource should be subjected to and more so. There are a number of guides on evaluation of Web resources. Matthew Ciolek (1996) and Hope Tillman (2000) have produced excellent Web resource guidance. Stability or permanence is cited as one key criterion for collection selection. There is growing evidence that Web site and page stability can be predicted with some degree of probability (see Koehler 2002, Bar-Ilan and Peritz 1999).

There are Web resources that meet quality and stability tests. Some of these are the e-journals that have begun to proliferate. University and government based Web sites provide extensive information. Many government agencies are migrating publicly accessible documents from paper to electronic format. As these trends increase libraries will have to consider these Web resources as appropriate for collection.

The Web may also serve as a viable substitute for expensive online databases and some traditional collections. Susan Lewis-Somers (2001) has found that there are a number of legitimate high quality free online resources that can be used in place of Westlaw or Lexis for legal research. Indeed, one can draw on *Thomas* and a multiplicity of other government e-publications and services to meet niche requirements.

The Web is sometimes seen as a "free" resource that can be "incorporated" into library collections without regard to collection objectives. In some sense, the Web is a free resource, but the process of evaluation, incorporation, and maintenance of Web resources in a collection is complex and expensive.

Finally, Keller (1992a) makes the very important observation that despite the metadata and sophisticated access systems, access is for the most part a question of "to what" rather than "how." Libraries must maintain good, current, and appropriate collections – however constructed – to meet the needs of their users. In order to maintain those collections, Peggy Johnson (1997) argues persuasively for formal electronic collection policies that reflect the changing landscape and that provide information workers with the guidance and a decision framework. As we have seen, in an Internet world, the need to discover and select quality relevant materials is as important as in the "print" world but it is as yet very difficult to do. Guidelines, such as those provided by Tillman (2000) or Ciolek (1996) are critical and should be employed with rigor.

2.4 Acquisition

In a paper-based world, the intellectual property had a physical form and, therefore, only one person could possess or use it at a time. Acquisition was the purchase of a commodity. Certainly, with the advent of photocopying machines, issues arose on the premise of one copy, one possession. To deal with this, copyright

has traditionally been balanced by the fair use doctrine. Copyright and fair use is discussed in more detail in another paper in this lecture.

Because of the economic implications for these intellectual property issues in the digital environment, in recent years, information suppliers have begun to move away from the sale of information to the licensing of information. The digital revolution has significantly changed the ways in which information can be packaged. It need no longer be offered within "physical packages." It can now be transmitted from producer to publisher to end user electronically. That may well render the principle of first sale moot because information containers need not be used. It also means that limits are placed on the ability of the licensee to transfer or transmit information to third parties. Under most licenses, lending practices and in some cases universal access is either prohibited or restricted.

With this background in mind, organizations may follow an ownership or an access path. Whichever they choose, they must develop acquisitions policies. Acquisitions of electronic or digital materials often entail a set of decisions that differ somewhat from paper (Pinfield 2001). Some of these decisions entail organizational questions. Are electronic acquisitions treated as an intrinsic part of the library collection or are they categorized as "other resources?"

As Stephen Pinfield has shown (2001), electronic acquisitions are not without their costs. These costs include the cost of the document (usually in the form of license). These services are offered using a variety of pricing models – individual subscriptions, bundled subscriptions, joint print and e-journal subscriptions, maximum number of users, and so on. The negotiations between library purchaser and licensor vendor can be time-consuming and complex.

2.5 Access Agreements

Now that acquisitions have taken the form of licenses rather than purchases, there are critical differences between traditional and electronic access agreements. Licenses represent permission or authorization for one party to use the property of another under a prescribed set of conditions. These licenses may limit the number of users for a database at any given time, they may limit the range of authorized users, and they place a temporal limit on access. Typically, full text providers, particularly e-book databases, place a limit on the number of users at any given time. This is analogous, they argue, to the traditional model. There can be only as many users of a print book as there are copies of the book at any given time. Harris (2002, p.100) also makes important distinctions between the sale, the assignment, and the licensing of rights. An assignment of rights, unlike a sale of rights, is the non-exclusive permanent transfer of rights of access to the item under consideration. Licenses are less permanent and are analogous to renting rather than buying the object. A license or an assignment may specify the conditions under which copyrighted material may or may not be used.

Many national and international associations have developed guidance to help information managers negotiate licenses or to better understand their implications for libraries and other users (for example, the American Library Association, the American Association of Law Libraries, the Association of Research Libraries, the Australian Library and Information Association, the Canadian Library Association, the Colegio de Bibliotecarios de Chile, the European Bureau of Library, Information and Documentation Associations, to list a few.) These concerns have sometimes resulted in a number of very specific agreements, as for example the International Federation of Reproduction Rights Organizations (IFRRO) – International Group of Scientific, Technical and Medical Publishers (STM) Joint Statements on Electronic Storage of STM material of 1992 (http://www.ifrro.org/papers/stmjoint.html) and 1998 (http://www.ifrro.org/papers/stmjoint2.html).

Lesley Ellen Harris (2002) has prepared a guide to digital licensing under the American Library Association imprint. Her book provides a series of checklists and commentary. For example, she advises her readers to avoid verbal agreements in favor of the written; in part, because of the potential for future misunderstanding (Harris 2002, xv). Special issues arise when negotiating licenses for library consortia, so the International Coalition of Library Consortia (ICOLC) provides special guidance in this area.

There are also a number of symposia and other training offered on digital copyright and associated issues. Some examples include the International Summer School on the Digital Library in the Netherlands (http://www.ticer.nl/index.htm) sponsored by the Tilburg International Center for Electronic Resources at Tilburg University and the Libraries in the Digital Age Conferences held annually in Dubrovnik, Croatia (http://knjiga.pedos.hr/lida/).

In addition, many professional organizations provide guidance on access, digital content, copyright, and related issues. Examples include:

- Association of Research Libraries (http://www.arl.org/scomm/licensing/)
- Copyrightlaws.com (http://www.copyrightlaws.com/index2.html)
- International Coalition of Library Consortia (http://www.library.yale.edu/consortia/)
- International Federation of Library Associations and Institutions, Licensing Principles (2000) (http://www.ifla.org/V/ebpb/copy.htm)
- Stanford University Libraries, Copyright & Fair Use (http://fairuse.stanford.edu/)
- Yale Library, Licensing Digital Information: A Resource for Librarians (http://www.library.yale.edu/~llicense/)

3.0 ELECTRONIC INFORMATION SERVICES

The World Wide Web is a complex information medium. It is both a repository for information and a transmission vehicle. It provides free public access and increasingly fee-based access to an immense body of digital material. The Web also supports a wide range of interactive services including banking and securities trading. E-commerce has moved into many other areas and it is now possible to purchase a wide variety of goods and services on line. Over the last several years, countries such as the United Kingdom, Canada, Australia, the United States and Lithuania are using the Web to disseminate information and to provide online services from government to citizen, government to government and between agencies of the government.

The advent of electronic information services has created a new set of demands for information providers. These services include new reference models, new means for information discovery and delivery, and demands for user and personnel education in the uses of the new resources and technologies. It has also prompted a re-examination of the rights and responsibilities of information providers, intermediaries, and end users (see, for example, American Library Association 2000).

A number of services are now offered online that, heretofore, were provided in person or through other print means. Online includes electronic reference and electronic document delivery systems. These services have been expanded to include automated information delivery and built according to various interoperable standards. Electronic information services that have been created include interactive e-commerce and e-governance services as well as various organizational database management needs (including registrations, membership renewals) and other functions.

The advent of electronic information services has also prompted new interest in artificial intelligence systems or agents to facilitate the delivery of information services. These range from natural language processing (see Jacquemin 2001) to the creation of content (see Bringsjord and Ferrucci 1999). These are future directions for services and are only mentioned here.

3.1 Electronic Reference

Electronic reference has come to mean several different things. By one definition, electronic reference is interpersonal reference information management using electronic means for the patron query and for the reference response. Libraries have employed this model using telephones for years. E-mail, instant messaging and chat have added a new dimension to the reference relationship. In this form there is still a one-to-one patron to librarian exchange.

An Association of Research Library (ARL) survey found that in the 10 years from 1991 to 2001, the median number of reference questions asked at reference desks dropped by nearly 30,000 (Ronan and Turner 2003). Research libraries responded to this phenomenon by locating a librarian at the point of interaction between their users and their online resources by implementing chat also called digital, real-time or live reference services. Libraries have formed partnerships to provide 24/7 access to chat reference service for their users. The Boston Library Consortium Ask 24/7 (http://library.brandeis.edu/247/) and Maryland Ask Us Now (http://www.askusnow.info/) are examples.

Chat Reference: A Guide to Live Virtual Reference Services (Ronan 2003) and the ARL Chat Reference Spec Kit (2003) are both useful guides designed to aid librarians in the implementation of this service. The sources provide information on selecting software, training staff, and evaluating the service. An executive summary of the Spec Kit can be found at http://www.arl.org/spec/273sum.html.

The User Group for Questionpoint, chat and email reference service software developed by the Library of Congress and OCLC, has created a living digital reference document (http://www.loc.gov/rr/digiref/QP_best_practices.pdf) for the purpose of supporting digital reference services as they evolve (QuestionPoint 2003). Policies and best practices for digital services are discussed.

Libraries are currently conducting studies to evaluate and assess the effectiveness of the chat reference in answering reference questions. In a pilot study conducted by the University of Maryland (http://www.dlib.org/dlib/february03/white/02white.html), the researchers found that librarians answered questions with a high level of accuracy but when it came to escalating a question to proceed deeper into the research process with the questioner, on the whole, librarians were less than adequate in this regard. The finding suggests that the interactive aspects of chat reference need to be further developed (White, Abels, and Kaske 2003).

A second model for e-reference has been developed and is more impersonal. Often, through email or Web-based queries, patrons place reference questions to anonymous reference librarians. The Internet Public Library provides an “Ask A Question” box with a pledge to respond within three days (http://www.ipl.org/div/askus/) The British “Ask-A-Librarian” service (http://www.ask-a-librarian.org.uk/) incorporates local reference librarians in an online service, with a pledge to respond within two working days. The Virtual Reference Desk (http://www.vrd.org) infrastructure developed by the Information Institute of Syracuse University and a network of partners provides software that triages reference questions through a series of experts who are linked by the network. Participating sites can be found through the AskA+ Locator which is organized by subject.

The National Information Standards Organization (NISO) in the US recently released a draft standard, the Question & Answer Transaction Protocol (NISO, 2004). The protocol defines a method and structure for exchanging data between digital reference services. The draft standard is available for trial use by implementers until April 2005 when it will be made available for comment as a draft NISO standard.

A third model establishes a reference-like interface, the "electronic reference desk." Libraries provide information portals with selected useful online information finding tools. Examples are given in Table 1.

Table 1: Examples of Electronic Reference Desks

Reference Service Title	Location	URL
The Virtual Reference Desk	Wageningen UR Library	http://library.wur.nl/desktop/vrd/
Electronic Reference Resources	Sourasky Central Library	http://www.tau.ac.il/cenlib/reframe.htm
Electronic Reference Shelf	McGill University	http://www.library.mcgill.ca/refshelf/swsindex.htm
The Online Library	University of London	http://www.external.ull.ac.uk/ref.asp
Oxford Reference Online (fee-based subscription)	Oxford University Press	http://www.oxfordreference.com
Ready Reference Shelf	University of Michigan	http://www.lib.umich.edu/refshelf/
Virtual Reference Shelf	NASA Goddard Space Flight Center Library	http://library/vrs/vrs.htm

Libraries have also begun to place online pathfinders on their Web sites. Pathfinders are pre-prepared reference tools designed to point users to resources for commonly asked reference questions. These pathfinders may focus on locally held resources or they may take advantage of Web-based resources. For an example of a set of pathfinders, see the Internet Public Library pathfinder page at: http://www.ipl.org/div/pf/. Building pathfinders requires some html knowledge, excellent reference skills, and a grounding in Web evaluation techniques.

FAQs (frequently asked questions) are a variant on pathfinders and "ready reference" that have developed in the Web environment. FAQs, as the name implies, are lists of questions and answers to those questions client, patrons, browsers, and others have asked of the Web site creator. They provide a ready resource to many questions that might be asked and save time for all parties.

Increasingly e-reference has begun to blend these techniques together. A true reference site may combine pre-set access to a Reference Shelf, access to specific librarians via e-mail, and a network of additional experts that provide support for more detailed or tertiary questions. FAQs and pathfinders may also be provided, and they may be built automatically as a result of collecting the answers to previously asked e-reference questions. An example is the Virtual Reference Desk at Wageningen UR Library which includes an extensive reference shelf, a search engine, and an e-mail connection to a librarian. Additionally, many librarians are using the transcripts from chat and email sessions to create searchable knowledge bases that capture the knowledge of librarians. The New York Public Library's "Ask A Question" service (http://ask.nypl.org/) includes chat reference in both English and Spanish, email reference, access to telephone reference, and a Q & A archive, a searchable knowledge base.

3.2 Information Discovery

The Web was once perceived not only as a supplement for libraries but as a replacement for them. Perhaps Louise Addis is the first information professional to appreciate the opportunities the Web can offer as a transfer medium (Berners-Lee 1999: 45) to support a variety of disciplines and library needs (Henderson 2000). It has been fairly well demonstrated that the Web is neither a library nor a substitute for libraries (Koehler 1999). It is one of many resources in the information environment.

Therefore, one of the most critical issues for electronic collection managers is the heterogeneous nature of information resources and the proliferation of information discovery tools. While some portion of the Web content is indexed by Google, Yahoo and other Web search engines, these search engines do not normally provide access to the "deep or hidden Web", that part of the Web that is hidden in databases, is password protected or behind firewalls. In addition, many different discovery mechanisms exist when trying to integrate external resources with internal databases, web pages or documents in document management systems. New metasearch tools, also known as federated, broadcast, and cross-database search tools, are being deployed to allow these disparate, heterogeneous resources to be searched simultaneously and then displayed from their native systems in a single interface. Examples include the Google Appliance, Goldfire, WebFeat, Vivisimo, ExplorIt from Deep Web Technologies and portal products such as Autonomy. A listing of metasearch and metacrawler engines is available from Search Engine Watch (http://searchenginewatch.com/links/article.php/2156241#reviews) along with links to related articles.

The metasearch or federated search tools are still being perfected with much debate surrounding their use. Encouragingly, as the federated search tools have evolved, services such as authentication, merging and duplicate identification have been added. However, problems still remain regarding relevancy, institutional repositories, and the one-size-fits-all philosophy that is behind metasearching. The ability for a cross-database search tool to return results from several sources is possible but relevancy ranking the items that are returned is difficult. Complications arise when library catalogs (MARC records) or sources that use OpenURL-based systems to link to full text articles are used. Setting the algorithm for the de-duplication of records located in several databases so that duplicates are eliminated but items that are not really duplicates are included is a challenging task (Tennant 2003b).

For institutional repositories to be included in a federated search process, libraries would have to harvest the metadata from the repositories and then make the harvested data a target, a database in metasearching, which could be searched by the metasearch tool (Tennant 2003b). The Open Archives Initiative (OAI) protocol for metadata harvesting (http://www.openarchives.org) was formed to encourage the creation of e-print repositories and to facilitate research information distribution. Several cross-database science and technology search tools have been established using the OAI protocol. Arc, (http://arc.cs.odu.edu) a cross archive search service, is one such tool, and, as of September 1, 2003, it contained 6,475,000 records from 160 repositories (McKiernan 2003). For more information on metasearching of institutional repositories follow the three part series on open archives initiative service providers in *Library High Tech News*.

At this time, federated search engines serve some user groups better than others. Undergraduate students are well served by federated search tools as they are looking for general or introductory level information on a given topic. Graduate students and faculty need thorough coverage within a given discipline and need the advanced searching capabilities of the native subject database (Tennant 2003b). Further, database producers feel that federated searching loses the search efficiencies unique to their native databases. For example, when metasearching JSTOR, it is not possible to search within a discipline as it is intended in the native design leaving the search result not subject focused. Licenses with database providers may or may not allow for the

inclusion of their databases in a federated search. When licenses are renewed, federated searching will certainly be addressed (JSTOR 2004).

3.3 Information Delivery

Libraries and other information providers are moving to augment or change traditional models by providing a wide array of electronic services. A well known example is the library at Los Alamos National Laboratory. This library is helping to meet the information needs of LANL scientists and engineers with the Library Without Walls concept. The LANL LWW provides services to its patrons at all times using electronic information delivery, enhanced data base access, and customized linking between bibliographic and full text resources. It also provides electronic information dissemination through a "MyLibrary" service to its patrons.

Selective Dissemination of Information (SDI) and document delivery systems are of long standing in the library community. Document delivery is a library-managed courier service to move requested documents from the repository to the end user and back. Many libraries have long provided such services by moving physical objects. One of the largest such delivery services is the British Library's Document Supply Center (BLDSC). In recent years, the BLDSC has moved toward extensive electronic document delivery by fax and through transmission of digital documents by ftp, telnet, e-mail attachment or the Web. Similarly, the services offered by such companies as Amazon.com, Borders, or Barnes and Noble represent a form of document delivery provided by the commercial sector.

SDI represents a slight variation on the document delivery model. Under the SDI model, documents are delivered to end users based on some criteria other than specific demand for the object. This may be a user profile developed by the librarian in cooperation with the end user based on end user interests. Many vendors and some libraries suggest additional documents of potential interest to the end user by offering "more like these" services.

Electronic Information Delivery (EID) is a variation on SDI models that uses the growing power of Web technology and content-oriented standards to respond to user requests from distributed content sources. A number of digital EID systems have been developed based on eXtensible Markup Language (XML), the building of complex indexes, and filtering mechanisms (Altinel and Franklin 2000). XML is designed specifically to structure information so that user queries can get better content responses. It uses "HTML-like" syntax to provide application specific meaning to digital documents through character string mark up. A related technology is RSS (Rich Site Summaries), a series of XML-based formats for the syndication of news and news-like items. It allows "news aggregator" programs to monitor these feeds and to respond when changes or updates are identified. Libraries and other information aggregators can "capture" these feeds and present them in their own interface formats. The use of RSS and "news aggregators" is especially popular in the web blog community.

These SDI "push" technologies automatically provide end users with information based on predetermined interests. These interests may be established from a profile developed by interaction with the end user, or they may be developed based on the behaviors over time of the end user. News, weather bulletins, and stock quotes are common examples of information pushed or streamed to the desktops or pages of information consumers (Dysart and Jones 1995).

"Pull" or demand technologies require the user to be proactive. They require interaction at the transaction time between the end user and the information provider. This is, in classic terms, the interaction between a reference librarian and a patron (Small Helfer 1997). More recently, interactions between end users and search

engines represent "pull" interaction because the end user is involved in the identification and "pulling" to himself or herself of the desired documents.

Most hybrid libraries offer some form of electronic pull. Clearly, the OPAC represents an interactive system for information discovery and retrieval. Some academic libraries offer electronic reserve services for their faculty and students. Reserve librarians place scanned or digitized documents in the service to be retrieved and viewed remotely by students on demand. Full text databases, like netLibrary, ingenta.com and OCLC's ArticleFirst may be used to pull documents either directly by the end user or through a librarian intermediary.

It is most interesting to note that the relationship between reference and information delivery is rapidly being redefined. Whereas information identification and then full text delivery used to be two distinct processes, as information is born and managed digitally, the identification of what a user needs is often only a click away from that information.

4.0 USER AND PERSONNEL EDUCATION

User and personnel education in information resources and access have been traditional library functions from training in school and public libraries to orientation and bibliographic instruction at the university level. Because of the explosion in digital and online resources and the frequent change in technologies and standards, libraries have had to develop in-house training programs for users and staff. They have also brought pressure on library schools to enhance the skills of their graduates in these areas. In addition, the advent of the Web made easy access to information over the Internet a reality. With the advance of search engine technology, came a revolution in information literacy and information use. This brought new demands on information professionals as well as many new public policy issues such as the digital divide.

The digital divide and ways to democratize information and technologies were key issues in the discussions of the World Summit on the Information Society held in 2004. Initially sponsored by telecommunications companies to promote wider access to undeveloped markets, the social and educational component of this global forum quickly became apparent. The outcome includes statements regarding the need for education and access in order to reduce the digital divide as much as possible.

In many countries, academic and public libraries have been identified as key institutions to assist in bridging the digital divide. In order to address these issues, they must train and retrain staff in the use of electronic technology and familiarize their patron or client base in the use and scope of those technologies.

In the United Kingdom, The Peoples' Network is a government funded undertaking to bridge the digital divide. Using lottery-derived funds passed through the New Opportunities Fund, the object of the project is to provide universal Internet access, digitize local resources, and teach online skills to the public. The program offers a useful checklist for personnel training and the areas that should be covered. To help implement the plan, public library staff are to be provided with an eight-point set of Information and Communications Technology (ICT) skills (http://www.peoplesnetwork.gov.uk/training/background.asp):

"1) A grounding in core ICT fundamentals;

2) Understanding how ICT can support library staff in their work;

3) Health and safety and legal issues in the context of ICT;

4) Knowing how to find things out on behalf of users;

5) Using ICT to support reader development activities;

6) Using ICT to support users to ensure effective learning;

7) Ensuring effective management of ICT resources in libraries;

8) Knowing how to use ICT to improve their own professional efficiency and to reduce administrative an bureaucratic burdens."

More advanced skills include:

"1) Net navigator – in-depth searching skills; validating Web sites; and using alerting services;

2) Information technology gatekeeper – web design skills; mounting and updating information; setting up and managing email databases; designing specialist interfaces; and setting up digital links;

3) Information consultant – analysis and diagnosis of users needs; awareness of information sources; building partnerships with other information providers; and information design and presentation;

4) Information manager – strategic planning; understanding regulatory and legislative requirements; content creation skills; and

5) Educator – training other staff and users to use ICT effectively and designing learning materials and programmes."

Most libraries have patron training and education programs on "how to use the library." This may involve in-library training and tours and, sometimes, online tutorials. Most of these programs now include the use of electronic resources. Students consistently show poor quality selection ability when choosing Web materials. If Web documents are to be incorporated in library collections, it is incumbent on libraries to provide assessment as well as search/retrieval training to patrons as well as staff. Kathy Schrock's Guide for Educators on Critical Evaluation Information (http://school.discovery.com/schrockguide/eval.html) provides survey forms and other information on teaching students of different ages to perform this critical evaluation.

5.0 CONCLUSIONS

Electronic collection management and electronic information services are in a period of rapid transition. Information organizations are undergoing redefinition. New forms of digital libraries and information collections are providing more information to more users more easily and on demand. These changes are being felt and responses are being made by information professionals throughout the world. The value of information is more appreciated than ever. Information collections are no longer geographically bound. Using Web access, it is possible to search the OPACs of many of the world's libraries and online resources from major primary and secondary publishers. Thus online and hybrid libraries have global reach. With global reach comes global responsibility.

The technology used to manage the information changes allows for extensive innovation in information selection description, distribution, retrieval, and use. The new e-publishing environment requires new ways to assess information for the purpose of selection. There is a new array of information markup and cataloging systems for collection management that, in turn, supports an equally growing array of information services for information producers, consumers, and intermediaries.

The full story for electronic collection management and electronic information services has yet to be told. These many changes and challenges give new meaning to the expression "may you live in interesting times."[2] We are indeed living in interesting times and they will become more interesting still.

6.0 REFERENCES

Altinel, M. and Franklin, M. (2000). Efficient filtering of XML documents for selective dissemination of information. *VLDB 2000: Proceedings of the 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*, pg. 53-64. [Online]. Available: http://www.informatik.uni-trier.de/~ley/db/conf/vldb/AltinelF00.html [24 June 2004].

American Library Association. (last updated 10 October 2000). Access to electronic information, services, and networks: An interpretation of the Library Bill of Rights. [Online]. Available: http://www.ala.org/alaorg/oif/electacc.html [21 April 2004].

Arms, W. (2000). *Digital libraries*. Cambridge: MIT Press.

Atkinson, R. (1993). Networks, hypertext, and academic information services: Some longer-range implications. *College & Research Libraries,* 54: 199-215.

Bar-Ilan, J. and Peritz, B.C. (1999) The life span of a specific topic on the Web: The case of 'Informatics': a quantitative analysis. *Scientometrics,* 46(3): 371-82.

Berners-Lee, T. (1999). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor.* New York: Harper.

Billings, H. (2000). Shared collection building: Constructing the 21st Century relational library. *Journal of Library Administration,* 31(2): 3-14.

Birdsall, W. (1994). *The myth of the electronic library: Information management and social change in America.* Westport, CT: Greenwood.

Bjoernshauge, L. (1999). Consortia building and electronic licensing as vehicles for re-engineering academic library services: The case of the Technical Knowledge Center and Library of Denmark (DTV). Issues in Science and Technology Librarianship. [Online]. Available: http://www.library.ucsb.edu/istl/99-spring/article5.html [21 April 2004].

Bringsjord, S. and Ferrucci, D. (1999) *Artificial intelligence and literary creativity: Inside the mind of Brutus, a storytelling machine*. Mahwah, NJ: Lawrence Erlbaum.

Buckland, M. (1997, orig. pub 1992). Redesigning library services: a manifesto. [Online]. Available: http://sunsite.berkeley.edu/Literature/Library/Redesigning/html.html

Budd, J. and Harloe, B. (1994). Collection development and scholarly communications in the era of electronic access. *Journal of Academic Information Management,* 20(5): 83-87.

[2] For an interesting discussion of the origin of the expression and the Chinese source urban legend for "May you live in interesting times," see Stephen E. DeLong. "Sidebar: Get a(n interesting) Life!" http://hawk.fab2.albany.edu/sidebar/sidebar.htm

Budd, J. and Harloe, B. (1997). The future for collection management, in G.E. Gorman and Ruth Miller, eds., *Collection management for the 21st Century: A handbook for librarians* (3-23). Westport, CT: Greenwood.

Bush, V. (1945). As we may think. *The Atlantic Monthly,* 176(1): 101-108.

Calhoun, K. and Riemer, J.J., eds. (2001) *CORC: New tools and possibilities for cooperative electronic resource description.* New York: Haworth.

Chen, C.C. (1998). Global digital library: Can the technology have nots claim a place in cyberspace? In Ching-chih Chen, ed., *Proceedings NIT '98: 10th International Conference New Information Technology, Hanoi, Vietnam, March 24-26, 1998* (9-18). West Newton, MA: MicroUse Information.

Ciolek, T.M. (1996). The six quests for the electronic grail: Current approaches to information quality in WWW resources. *Revue Informatique et Statistique dans les Sciences Humaines* (1-4): 45-71. [Online]. Available: http://www.ciolek.com/PAPERS/six-quests1996.html [21 April 2004].

de Sola Pool, I. (1983).Tracking the flow of information. *Science,* 221(4611): 609-13.

Dowlin, K. (1984). *The electronic library*. New York: Neal-Schuman.

Dysart, J.I., and Jones, R.J. (1995 January). Tools for the future: Recreating or 'renovating' information services using new technologies. *Computers in Libraries,* p.16+.

Ferguson, A. and Kehoe, K. (1993). Access vs. ownership: what is most cost-effective in the sciences. *Journal of Library Administration*, 19(2): 89-99.

Gessesse, K. (2000). Collection development and management in the twenty-first century with special reference to academic libraries: an overview. *Library Management,* 21(7): 365-72.

Harris, L.E. (2002) *Licensing digital content*. Chicago: American Library Association.

Harter, S. and Kister, K. (1981) Online encyclopedias: the potential. *Library Journal,* 106(15).

Henderson, M. (2000) FM Interview: Louise Addis. *FirstMonday 5* (5): [Online]. Available: http://www.firstmonday.dk/issues/issue5_5/addis/ [21 April 2004].

HyLiFe Hybrid Library Toolkit. (2002). [Online]. Available: http://hylife.unn.ac.uk/toolkit/ [21 April 2004].

Jackson, M. (1998). Maximizing access, minimizing cost: The Association of Research Libraries North American Interlibrary Loan and Document Delivery (NAILDD) Project: A five year status report. Access & Technology Program/NAILDD Project. [Online]. Available: http://www.arl.org/access/naildd/overview/statrep/statrep-9801.shtml [21 April 2004].

Jacquemin, C. (2001). *Spotting and discovering terms through natural language processing.* Cambridge, MA: The MIT Press.

Johnson, P. (1997). Collection development policies and electronic information policies. In G.E. Gorman and Ruth Miller, eds., *Collection management for the 21st Century: A handbook for librarians* (83-104). Westport, CT: Greenwood.

JSTOR. (2004) Metasearching JSTOR. *JSTORNEWS.* [Online]. Available: http://www.jstor.org/news/2004.02/metasearch.html [28 April 2004].

Keller, M. (1992a). Foreign acquisitions in North American research libraries. *FOCUS on the Center for Research Libraries, 12*(4), special insert.

Keller, M. (1992b). Moving toward concrete solutions based in fundamental values. *Journal of Academic Information Management* 18(3): 8.

Koehler, W. (1999). Digital libraries and World Wide Web sites and page persistence. *Information Research,* 4(4). [Online]. Available: http://InformationR.net/ir/4-4/paper60.html [21 April 2004].

Koehler, W. (2002). Web page change and persistence – A four-year longitudinal study. *Journal of the American Society for Information Science and Technology,* 53(2): 162-171.

Lancaster, F.W. and Warner, A. (2001). *Intelligent technologies in library and information service applications*. Medford, NJ: Information Today, Inc.

Lee, S. (2000). *Collection development for the electronic environment*. Binghampton, NY: Haworth.

Lee, S. (2002) *Electronic collection development: a practical guide*. New York: Neal Schuman.

Leggate, P. (1998). Acquiring electronic products in the hybrid library: prices, licenses, platforms and users. *Serials,* 11(2): 103-108.

Lewis-Somers, S. (2001). Electronic research beyond LEXIS-NEXIS and Westlaw: Lower cost alternatives. Gary Hill, Dennis Sears, and Lovisa Lyman, eds. *Teaching legal research and providing access to electronic resources*. New York: Haworth Press.

Lesk, M. (1997). *Practical digital libraries: Books, bytes, and bucks*. San Francisco: Morgan, Kaufman.

Lyman, P. and Varian, H.R. (2000). How much information? [Online]. Available: http://www.sims.berkeley.edu/research/projects/how-much-info/how-much-info.pdf [21 April 2004]. Summary: http://www.sims.berkeley.edu/research/projects/how-much-info/index.html [21 April 2004].

McKiernan, G. (2003). E-Profile: Open archives initiative service providers. Part 1: Science and Technology. *Library Hi Tech News.* 20(9): 30-38.

Mischo, W.H. (2001). Library portals, simultaneous search, and full-text linking technologies. *Science & Technology Libraries.* 20(2/3): 133-147.

Negroponte, N. (1995). *Being digital*. New York: Vintage Books.

NISO. (2004). Question/Answer Transaction Protocol. (Draft, April 2004). [Online]. Available: http://www.niso.org/committees/net-ref-protocol.html [13 June 2004].

Oppenheim, C. and Smithson, D. (1999). What is the hybrid library? *Journal of Information Science,* 25(2): 97-112.

Pinfield, S., Eaton, J., Edwards, C., Russell, R., Wissenburg, A. and Wynne, P. (1998 October). Realizing the hybrid library. *D-Lib Magazine* [Online]. Available: http://www.dlib.org/dlib/october98/10pinfield.html [21 April 2004].

Pinfield, S. (2001). Managing electronic library services: current issues in UK higher education institutions. Ariadne 29 [Online]. Available: http://www.ariadne.ac.uk/issue29/pinfield/ [21 April 2004].

QuestionPoint User Group. (2003) Library of Congress QuestionPoint User Guidelines. [Online]. Available: http://www.loc.gov/rr/digiref/QP_best_practices.pdf [28 April 2004].

Ronan, J. (2003) *Chat reference: a guide to live virtual reference services.* Westport, CT: Greenwood.

Ronan, J. and Turner, C. (2003) SPEC Kit 273, *Chat reference*. Washington DC: The Association of Research Libraries.

Small Helfer, D. (1997 May). Not your traditional librarian anymore! *Searcher*, p.66+.

Tennant, R. (2003a) Science Portals. *Library Journal*. 128(5): 34.

Tennant, R. (2003b). The right solution: Federated search tools. *Library Journal.* 128(11): 28-29.

Tillman, H. (2000). Evaluating quality on the Net. [Online]. Available: http://www.hopetillman.com/findqual.html [21 April 2004].

Wells, H.G. (1937). *World brain*. London: Methuen.

White, M., Abels, E. and Kaske, N. (2003). Evaluation of chat reference service quality. *D-Lib Magazine*. February 2003. [Online]. Available: http://www.dlib.org/dlib/february03/white/02white.html [21 June 2004].

Economics of Electronic Information Provision

Graham P. Cornish
Consultant, Copyright Circle
33, Mayfield Grove
HARROGATE
North Yorkshire HG1 5HD
UK

Graham@copyrightcircle.co.uk

ABSTRACT

The paper covers the economics of preparing and providing published information. The role of the different players in the publishing chain are examined including authors and their institutions, publishers, reviewers, editors, distributors and users (readers). The costs of acquiring manuscripts, peer review, editing text, and presenting it in an acceptable format will be outlined. The economics of marketing as an aspect of availability will be explored. Different markets will need to be identified and their differing economic relationships to the publisher and the user will be described. These will include libraries, institutions, individuals and groups such as clusters of learners. The role of so-called "grey" literature is emphasized throughout as an alternative model for making available scientific information in a non-commercial environment. Economic models for providing access will be analyzed including consortia agreements, different publisher models. Examples will be used from Emerald Press, Elsevier Science Publishing, JSTOR and others. The important work of the International Coalition of Library Consortia (ICOLC) will be emphasized as a useful model for others to follow. The many different contexts in which information has to be delivered will be studied against the models currently available. This will include distance learning, the client-server relationship and the problems of access to information in Third World Countries. The changing role of document delivery will be used to contextualize the issues of pay-per-view or pay-per-use and the whole complex issue of who pays will be put forward as a final challenge to participants.

1.0 INTRODUCTION TO PUBLISHING MODELS

Before exploring the specific issues surrounding electronic publishing it is crucial to examine the traditional patterns that have existed for some time in the paper publishing world as many of these practices have been carried over into the electronic context and newer models are emerging from these well-established patterns. There are basically three models – totally commercial; commercial but assisted by other factors, and "grey literature". The first two have much in common and will be treated together but grey literature is a different creature and needs to be mentioned repeatedly as the exception to the general rule.

Additionally it is valuable to be able to compare and contrast the past (and the present) with the present and the future as we are in a major time of transition. This is equally true for the impact of electronic media on publishing as it is on political changes on market economies and publishing and distribution mechanisms. In many areas the publishing industry has not caught up with technology. As John Harvey Jones, the former Chief Executive of the chemical firm ICI said:

> It is remarkable how little the publishing industry has changed, and how slow it has been to deal with the problems that have arisen for it over the years. The falling demand for books spells trouble, yet the number of titles published keeps going up. This is obviously in part due to falling production costs for books (Harvey-Jones 2002).

This is less true for the scientific journal publisher but, even then, some parallels can be drawn.

1.1. The Model for Scholarly Publishing

The word "scholarly" in this context does not relate to purely academic research as carried out in many universities or institutes of theoretical research. Rather it is used in the context of scientific and technical literature in its broadest interpretation. The models are different from those for leisure and recreational publishing. For scientific monographs the traditional model has been that primarily commercial publishers will adopt one or two definite strategies. Either they will establish themselves as the major publisher in a subject-related discipline or they will concentrate their marketing on a specific segment of the market such as a particular scientific community, geographical or linguistic area or economic group. An alternative model is that the publication is produced with a subvention of either the sponsoring institution or an independent trust or similar body with an interest in carrying forward the provision of information and research in the discipline concerned. This is particularly true of the university presses found in a considerable number of academic institutions which may be mirrored by similar presses in the various publishing programmes of the Academies of Science found in many countries. In these cases the primary concern is the publication of the information rather than the commercial viability of such a publication in its own right. A combination of these models can be found in presses which have both a non-commercially viable list and a highly "popular" one. Such an example is Oxford University Press in the UK which publishes very scholarly monographs with a very limited market (and no hope of commercial viability) and also a wide range of popular titles ranging from dictionaries and atlases to children's books and school textbooks. With a certain amount of cross-subsidy and support from the university itself OUP can manage to produce high-quality scholarly works in disciplines ranging from archaeology to zoology. This model is now re-emerging for journals as individuals and institutions are being asked to contribute to the cost of publishing their research.

1.2. Grey Literature

However, it must not be assumed that the commercial or scholarly model is the only one already in existence and vast amounts of scientific data are readily made available through non-commercial outlets. This is usually given the generic name "grey literature". People have often asked for a definition of grey literature and those who are used to handling it usually say "Hard to say but you know it when you see it." Essentially this term is used to describe that output of the scientific community which is carried out purely for the benefit of the community itself. It is not done for commercial gain and often costs the "publisher" quite a lot in printing and distribution terms. Material is often produced direct from camera-ready copy or even photocopied from an original. Increasingly it is stored on a PC and printed out or sent as an email attachment as required. It rarely enters to commercial marketing chain and is not available through normal bookselling or subscription agency outlets. It may not be controlled bibliographically through the normal channels although the efforts of EAGLE (the European Association for Grey Literature) in constructing a European-wide database has done something to correct this particular weakness. Grey literature may be in paper form or often in microform such as the huge output managed by NASA and ERIC in the USA or many nuclear and environmental agencies in Europe. The different model used here for publishing is exemplified by the fact that many of those who produce this material are happy for it to be photocopied for further use without any permission being sought or royalty paid in lieu of purchase. Grey literature falls outside the remit of most copyright licensing agencies and the licences they offer.

2.0 THE PROCESSES IN THE PUBLISHING CHAIN

2.1 The Role of the Author

It is a truism that publishers cannot publish unless someone creates something for them to publish! Therefore the process must begin with an individual or a group of individuals preparing a piece of work

which they wish to have published. The group may be a team of scientists or simply the management of an institution.

The motives for scientific and technical publishing are most often that the researcher wishes to make known the findings of the research carried out and also to enhance their national or international reputation in their field of expertise. This is particularly important in an increasing number of academic environments where awards from university authorities or central government to individual departments is dependent on the number of academic papers published by the staff of that department. This will be discussed further under "peer review". It is certainly true that a small minority of authors actually publish to make money. These are mostly authors of standard textbooks used by predominantly undergraduate students so that sales are often in bulk and guaranteed on a recurring cycle as they are recommended reading on university courses. In the case of grey literature concepts of enhancement of reputation and certainly commercial gain are entirely lacking.

2.2 Editors

Most publishers are well-versed in their own discipline – publishing. They rarely know for themselves what material is the most appropriate to publish or, indeed, even how to find it and certainly not how to evaluate it. Therefore publishers use specialists in the field of expertise to edit journals or series of monographs to ensure that the highest quality material is published under their imprint. Editors are usually drawn from the scientific community and have considerable knowledge of the latest developments in their field. They can evaluate a paper or text for a monograph from a technical as well as literary viewpoint and decide whether or not it is suitable for publication from the point of view of originality, content and relevance. As no editor can have a total overview of a subject most editors turn to a team of reviewers in the discipline concerned for advice and comment on papers submitted for publication. This is usually called "peer review". This process is, of course, entirely absent in grey literature as the decisions about "publishing" will relate only to whether or not the issuing institution wishes to disseminate the information or not.

Once material has been received and it is agreed it is appropriate for publication the editor must then decide how and when to publish it. Decisions about format, timing and context will all need careful consideration.

Having reached the point of deciding what to publish and when, the production process proper begins. Texts need to be prepared in a standard format and edited for linguistic correctness (something of a debating point in many scientific communities), form of references and bibliographical notes, accuracy of content and suitability of the format of diagrams, drawings and illustrations.

Once these points have been settled and the text is ready from a professional scientific point of view it must then be submitted to the publisher for printing. The way in which this is done will vary from one publisher to another. Again, grey literature avoids most of these bureaucratic structures but at the cost of usually having a rather boring appearance.

2.3 Publishing and Distribution

Once the editor has performed all the necessary tasks of deciding where and how a work will be published the publisher will then take over the technical processes of printing, binding and packaging for distribution to the public. Monographs and journals require very different distribution strategies as is fairly obvious.

How then are these different links in the publishing chain achieved in the traditional publishing model?

3.0 THE TRADITIONAL MODELS – THE ROLE OF THE AUTHOR

3.1 The Origins of Manuscripts and Texts

Different segments of the publishing industry deal with different models of creativity. In the case of the recreational market (novels, leisure magazines, for example) publishers will often commission works on a speculative basis that they will sell. First time authors will probably submit manuscripts proactively but, once established, publishers will contract with them to produce more novels. In the case of leisure magazines publishers rely almost entirely on identifying subjects they wish to cover and then commissioning a well-known author to write that subject up in a suitable format. In both cases the author will be paid although mechanisms for this will vary. This segment of the publishing industry is not one that will be discussed further in this chapter.

There are usually two complementary methods for editors to obtain papers for their journals or monographs for their series. Particularly in the former case authors may submit a paper for consideration proactively in the hope the editor will accept it and publish it. Alternatively the editor may identify a particular piece of research or a specific author who would enhance the reputation for the journal or publisher if included in the publications programme. In this case the editor will approach the persons concerned direct to try to persuade them to write an article or prepare a manuscript.

Authors come from many different backgrounds but essentially there are three major areas on which editors and publishers can draw. The academic world in terms of universities, higher education colleges and individual researchers (the honoured but much-neglected amateur) will provide a vast array of papers for almost any journal. The commercial and industrial sector of research will also be a rich ground to harvest.

These different sectors may have competing or conflicting reasons for wishing to have their papers published. Increasingly the boundaries between different types of organization are becoming blurred and it is sometimes impossible for an individual researcher to be sure whether they are working for a commercial company, a government agency or a university at any given moment.

These diverse motives and pressures lead to a series of problems for editors of academic journals but less so for those planning major monographs. The latter have a largely archival and teaching role in modern scientific and technical areas so pressures are more about kudos and possibly financial reward than those outlined below.

From an economic point of view it is rare (though not unknown) for scientific authors to be paid at all for publishing their papers in academic or scholarly journals. They may be paid royalties for monographs and these can either be a one-off payment or a percentage of the retail price for every volume sold. Putting aside the monograph element, this means that the publishing industry receives the raw data for its publishing programmes free of charge (although not free of cost). The costs for publishers are essentially in the processing of the articles themselves rather than in paying fairly small royalties to authors.

As can be seen the grey literature model is different here. There are no costs of peer review associated with authors or editors as material is often produced “in-house” within the research budget costs of the institution concerned.

3.2 Academic Authors

As academic institutions move from being funded entirely by the state (with contributions from students as appropriate) to having to fund many of their activities from other sources there is increasing need to build partnerships with commercial and industrial concerns to carry out for them on an agency or

partnership basis, some of the research which the company may wish to do but which it is more cost-effective to have done in a collaborative environment. On the other hand the issue of "publish or perish" is an important one for academics. Publish or perish essentially means that either a department (or individual author) must either seek publication of relevant papers in appropriate journals or else lose their status and funding. As mentioned above many departments now rely on their staff achieving publication in recognized journals or by well-established publishers in order to obtain funding for the department from either the central university or from national government.

Clearly editors will find themselves under considerable pressure from individual scholars and universities to publish papers by and from them to help bolster their reputation within the scientific world and also ensure adequate funding from public sources in the future. There may also be a further agenda that a particular researcher or university is anxious to secure a particular contract with an industrial partner and publication might be part of the strategy to obtain that contract.

Grey literature meets none of these criteria and so is not considered by academics generally as an available outlet for other than ephemeral research reports.

Generally speaking, academics are less interested in payment for publication than the other motives outlined above. However this is set to change in the electronic environment as will be explored later.

3.3 Industry and Commerce

Companies may wish to publish research data to promote their products and activities. There is also the incentive on some jurisdictions to provide evidence for a patent registration. Alternatively such companies may be reluctant to allow publication for reason of commercial confidentiality. These vested interests from commercial companies will necessarily transfer themselves to plans to ensure (or prevent) publication by an editor in either a particular journal or, indeed, any journal at all.

Most industrial companies have little interest in receiving payment for publication by members of their staff. The sums of money involved would be small in comparison with the total research budget and their motives lie elsewhere than in raising revenue. For this reason a considerable amount of grey literature emanates from industrial companies. It is cheap to produce and entirely within their control as to who obtains copies of it. Large stocks are not needed and the company, which is usually not geared up to meeting demands from the general public for its output, can concentrate on its real business whilst meeting any proven demand without undue costs.

3.4 Government Funded Organizations

The third possible major source of papers is the public sector where scientific and technical research is funded by government (local, regional or national) and there will be a desire to make findings assessable "for the public good". At the same time this same very public good may be the reason why some research data is once again withheld from publication. It may be sensitive in terms of defence or might be considered a potential source of public alarm by politicians who may find it more appropriate to keep the information for internal and restricted use only. Editors may find either that they are under pressure from government to become a further tool in their political plans or else that they are met with a stonewall when trying to obtain papers on a particular subject for inclusion in their journal.

Government is probably the greatest source of grey literature in the world. Government agencies have variable policies regarding the sale or acquisition of their publishing output, ranging from the almost unlimited access approach of the US government by virtue of the clause in the 1976 Copyright Act which prevents them from enforcing copyright in their publications (and therefore any subsequent use or republication of them) to the UK Government which, until very recently, took a highly protective

approach. Fortunately, in the latter case, this approach has now been relaxed and Her Majesty's Stationery Office (HMSO), as manager of copyright owned by the British crown, requires government departments to be much more open with their material including allowing it to be freely downloaded and copied. Further details of the UK government's open access policy can be found on their website at www.hmso.gov.uk.

Again, authors and institutions in this sector are less interested in payment than promoting their activities. Although increasingly the public sector in most countries is under pressure to earn revenue, this has not demonstrated itself as a major obstacle to getting material published in the past.

4.0 THE TRADITIONAL MODELS – PEER REVIEW

The peer review system is at the heart of academic publishing and is one of the most controversial elements in the electronic age. Essentially the system is a filtering mechanism to help editors and publishers to determine what should and should not be published. As mentioned above, editors will assemble team of experts in the field in which their journal is published and refer papers to them to consider if they are (a) suitable for publication as they stand or (b) essentially suitable but requiring further work or (c) unsuitable for publication in the journal concerned (or, perhaps, in any journal at all). The process is often carried out on a "double-blind" basis so that the author has no idea who the reviewer is and the reviewer is not given any information as to the identity of the author. Only the editor holds both parts of this information. This is intended to ensure that any personal prejudices or antipathy to a particular institution are removed from the vetting process and that papers are considered solely on their scientific and research merit. The editor is therefore put in a particularly delicate position when the different pressures on him or her are considered. The system may be used to ensure an impartial decision or it could be used as a shield behind which highly subjective judgments could be made.

The economics of peer review are often overlooked by those writing for or reading academic journals and texts. It is a very time-consuming and labour-intensive activity involving many hours of decision-making by editors and considerable administration to ensure it is carried out effectively. Papers are received by editors and usually acknowledged to make sure the author knows it has arrived and is being processed properly. Occasionally an editor will return a paper immediately as being unsuitable for the journal because of the subject matter. Journals whose content is devoted to management rarely want articles whose topic is ornithology (unless it is a paper on managing a bird garden in a zoo!). The editor will then identify a suitable reviewer and pass the paper on, keeping records of when and to whom the paper was sent. The reviewer must then find time to consider the paper and perhaps undertake some research to identify whether or not it is both accurate and current as to its information content. The paper is then returned to the editor who will then either decide to accept the reviewer's comments or possibly contact a second reviewer for a further opinion. Many editors actually pass papers to two or more reviewers immediately so as to obtain a balanced opinion. All this takes time which can slow down publication and which causes considerable irritation to many authors as they are seeking speedy exposure of their research to the scientific community. However, the costs of this are often borne by the very same group of institutions from which the authors themselves come. Reviewers are not paid for their time and effort and often squeeze in such activities between teaching or other research activities. Editors themselves rarely receive more than fairly nominal payment for their efforts.

Peer review is, of course, entirely absent from grey literature as the aims and objectives of the process are quite different. The decision as to whether to make a particular document available will rest more with a senior researcher, often personally involved in the actual research being described, or with a senior civil servant or politician in the case of government material.

5.0 THE TRADITIONAL MODELS – THE ROLE OF THE EDITOR

Texts may well pass backwards and forwards several times for revision and correction between editors and authors before a final version is agreed. Even when this is done using conventional electronic technology such as floppy disk or email it is still a costly and expensive process for both authors and editors. Editors will still be responsible for the overall layout as well as accuracy of the final text including the proper positioning and labelling of illustrations and correct structures for headings, sub-headings and footnotes. Whereas publishers used to accept text in a fairly "raw" state now they are unwilling to take "raw" text produced on a conventional typewriter (and even less in handwritten format) and material must be submitted in electronic format on disk or email with strict guidelines on format, layout and even coding.

Editors must submit text in a ready-to-print state and publishers no longer use sub-editors to carry out the finer points of house-style and bibliographic citation. This is increasingly becoming part of the editorial process. Such publishers as Emerald Press (formerly MCB University Press) transferred much of this work to the editor some time ago, requiring them to deliver text that was virtually ready for printing. Publishers increasingly take the view that their investment is in the packaging, marketing and distributing of products rather than their preparation. This, it is argued, is simply an extension of the trend towards camera-ready publishing which has been a feature of scientific publishing, especially for conference papers, for many years. It is worth noting that much of the camera-ready publishing was not carried out for strictly commercial purposes but rather as part of a larger research package to make material more widely available. Here quality of end-product was less important than the combination of speed and cheapness which this technique afforded. To move to something similar for strictly commercial publishing has caused some considerable debate in the academic community.

Once again, editing text is something which is usually done to grey literature with a light hand. Only essential corrections will be made and often there is no "house style" other than that reflected in general internal protocols which will be used to prepare the document in the first instance.

6.0 THE TRADITIONAL MODELS – THE ROLE OF THE PUBLISHER

Once material has been accepted, edited and prepared it must be published. This is a self-evident statement but one which needs careful consideration as the term "published" has many meanings and nuances. One useful definition which is used by UNESCO and other bodies is "issuing copies of the work to the public at a time when copies made in advance of receipt of orders are generally available to the public". In other words, a publisher produces printed copies which are usually put a warehouse and are therefore available to meet any orders as and when they come in. This clearly excludes any form of on-demand publishing such as holding a work on microfilm or a PC and printing off copies as and when the need arises or sending them by email as an attachment. However, to stay with this definition, it is clearly the primary activity of most commercial publishers and also of those who are working in a quasi-commercial market such as university publishers.

This is one of the most expensive elements of traditional publishing. Even by using modern printing methods a whole series of highly complex and skilled actions are now needed. Printing, collating and binding works is a high tech activity which may involve operators in two or three different countries or even continents. Material has to be shipped around from place to place in different formats and finally stored and dispatched to many different points often throughout the world. These points may be direct orders, booksellers, subscription agents or individual subscribers. The maintenance of subscription lists or delivery addresses as well as the whole complex area of billing, accounting and following up customers for payment is a huge charge on any publisher or supplier. This then links into the associated costs for the running of bookshops and subscription agents.

Essentially this is a model by which the publisher sells access to the information by producing multiple copies of the carrier of that information (usually a scientific journal or perhaps a monograph) which the customer then purchases and over which that customer has total control in terms of the physical documents. Naturally the intellectual property rights (usually copyright) are retained by other players in the information chain (discussed in more detail in another chapter) but the purchaser has the absolute right to do whatever they wish with the physical volume bought. This provides the purchaser with current access and an archival resource. Access is limited to one person per item by the physical nature of the material but unlimited access is possible in a sequential situation.

Once again, grey literature is outside this model. Few copies of reports in this category are produced and there is no concept of a market in advance of publication, if publication can be used in this context. There is virtually no warehousing problem and if there is real demand for a particular document then a visit to the company or departmental print room will usually ensure that enough copies can be made to keep everyone who needs one happy.

7.0 THE TRADITIONAL MODELS – MARKETING MATERIAL

It is clear that any publishing exercise requires marketing of one kind or another in order to achieve the required impact. For commercial publishers this is a vital element in any business plan. Failure to tell prospective customers what is available under what conditions and from where it can be acquired would negate the whole exercise. Therefore publishers expend considerable amounts of money, time and expertise in promoting their products. For those who have focused on a particular segment of the market as described earlier the challenge is less daunting as for those who have carved out a niche in a particular discipline but for those trying to expand their markets or penetrate new areas this can require a considerable outlay of capital. This may be true simply because particular monograph or new journal title crosses one or more disciplinary boundary and is a multi-faceted publication which should appeal to readers in a variety of areas. In addition publishers need to keep their products constantly before the potential market place. Monographs need to be advertised but also the subject of in-depth reviews in relevant journals. New and even established journal titles require review but also frequent citation in abstracting and indexing services and other bibliographical tools. It is for this reason that many publishers are happy to have their contents pages photocopied or digitized and used as current awareness bulletins by libraries and information providers. Although this may possibly stimulate photocopying for which the publisher may receive no direct compensation nevertheless the fact that the journal is repeatedly consulted and cited means that its value to the academic and research community is going to be considered high and therefore the journal subscription will not only continue but may become a multiple one. Given that cuts in library budgets are frequent and a global phenomenon it is essential that publishers can demonstrate the value of their products over against other titles, even if they are not in direct competition from the point of view of the subjects covered.

In the case of scientific and technical material the existence of the peer-reviewed journal is an essential element of marketing. The journal provides a ready-made package in which the consumer buys a range of pre-packaged products including technical articles, editorial comment, correspondence, advertisements and information about current and forthcoming events such as conferences and new projects. The journal has a reputation and the purchaser knows what is being offered in terms of quality of academic content, presentation, relevance and scientific integrity. The marketing of the learned journal is an essential elements in the economics of scholarly publishing as it provides a ready-made “label” recognized by the reader and the purchaser (the library or the institution) as well as by the author. It is this labelling that is essential for marketing but also for the status of the author. Publication in an esteemed journal ensures that status is either retained or enhanced in the scientific research community.

There is an additional facet to marketing which is often overlooked in the determined effort to adopt an economic interpretation of this term. As well as selling a product and generating revenue, marketing is

also about making a product known because that is the essential role of any publisher. Publishing is essentially making things available to the public whether for profit or not. Therefore every publisher has a role in disseminating information of various kinds. So whether or not sales are generated it is important that the product, and the information contained in the product, are widely known. This is equally true of the total commercial publisher and those whose commercialization is underpinned by other elements such as outside trusts or university funding. In fact many of the latter would put dissemination of information as a top priority beyond mere revenue generation. In the UK members of the Association of Learned and Professional Society Publishers (ALPSP) repeatedly make this point when discussing issues such as copyright, ownership of rights in journal articles and payment for authors. See their website at www.alpsp.org.uk and the journal *Learned Publishing*.

Once again grey literature makes little effort to market itself in the traditionally-received understanding of that term. However, like other publishers the producers of grey literature do want their work to be read and disseminated. However there will be a disinclination to promote the publication too vigorously in case demand is stimulated beyond what the producing organization can reasonably meet. Therefore "marketing" of grey literature will be more focused on making the document known to those who need to know about it rather than reaching out to those on the fringe or the merely curious. The use of bulletins and abstracting and indexing services will be more prominent here although some government departments may use the propaganda machinery of government and politics to make certain documents widely known and their availability easily understood. These factors will be influenced as much by politics as by economic aspects of information provision.

8.0 LIBRARIES AND ACCESS

In the past access has been via the published journal or, less important in volume terms but crucial in such disciplines as high-energy physics, the distribution of preprints. Rarely do individual scientists subscribe to journals because of their high cost; occasionally a researcher may build subscription to a journal into a project proposal but essentially access has been, and in many cases still is, via the library. The library is the knowledge storehouse of any institution and its very nature places it in a pivotal position for the dissemination of knowledge.

Traditionally librarians have collected and conserved material and, more recently, worked to make it more accessible and publicize what they have. However changing patterns of information provision could mean that users go direct to sources of supply without consulting the library at all. In theory this could mean the end of libraries in as we understand them today. It is most unlikely that this will happen because (a) libraries are themselves major sources of information provision and (b) users cannot have access to every source of supply and need guidance on what is the best and most appropriate source for their needs. As has happened in the case of databases there will always be a need for an intermediary although the role for that intermediary will change but not disappear. Nevertheless the role of the library will change from supplying information and documents and leaving the user to decide what is relevant or most interesting, to supplying packages of information, much of which has already been evaluated to reduce, if not eliminate, the "noise" factor. The desire, never mind the need, for information is a constant feature of current cultural patterns, particularly in the industrialized world. The information may be supplied in various ways: newspapers, journals and books, broadcasts, television, teletext, sound, video, images or online systems. These are all materials and carriers which are commonly found in libraries and handled by librarians. In the more sophisticated reaches of the information supply industry librarians are not simply renamed "information scientists" but transmogrified into "knowledge scientists". A knowledge scientist is not expected to provide information but to interpret it for the customer. This particular trend leads those in this situation to receive requests for appropriate data, suitably packaged, on a given topic or aspect of a topic. The resulting package may be a concoction of statistics, manipulated data, law, company information, economic projections and predictions and some documents. The knowledge scientist will be required to obtain such information, whether in the form of documents or other carriers, either locally or

from remote sources. The customer in this situation has little interest in where or how the document was procured so long as it supplies the needs of the time. Although this is at one extreme end of the information supply spectrum, it is nevertheless symptomatic of an increasing trend in the information industry at all levels. Documents are seen as vehicles for information in its widest sense. And "information" should not be understood in too narrow a sense. The content of a well established piece of non factual writing is viewed by many as a piece of information and the format in which it is delivered is far less relevant than the delivery itself. The growth in the use of audiocassettes by car drivers, by users of public transport or even those taking physical recreation to replace reading habits is a clear sign of this attitude.

8.1 Libraries are not Supermarkets

Many people have argued that the Library is a sort of information supermarket but this view is quite inadequate. Certainly supermarkets do collect a wide range of products, often competing with one another and therefore allow the customer to choose the products they require. It may also expose the customer to products and services of which the customer was previously unaware and had never thought of purchasing. Although the supermarket collects together a wide range of products the scope of this is dictated entirely by commercial requirements. Although some items will be offered for sale which are unlikely to generate much, or any, profit these are seen in an overall context of drawing customers in to buy other products which generate considerable income. The practice of deliberately offering one or two products at vastly reduced costs is a clear example of this. Nevertheless, overall, the purchasing policy of the supermarket will be dictated by commercial needs. In the same way there will be certain groups of customers which the supermarket will wish to discourage. For example, households consisting of single persons or elderly persons on reduced incomes may not be welcome in some shops. Their product requirements are often small and limited and unlikely to contribute to the overall income generation policy of the store. At the other end of the spectrum there will be high quality, high price products which would probably move very slowly from a supermarket environment and these would be excluded and therefore certain groups of customers who expect to find this sort of product readily available.

A library, on the other hand, will make strenuous efforts to collect all such materials it considers likely to be required by its users now or in the foreseeable future. There will be no consideration of commercial benefit but only the likelihood of meeting the requirements of the user community. Naturally acquisition is limited by budgetary considerations and judgments will have to be made about which services or information products to acquire. This will be governed by use and usefulness rather than income generation. At the same time no library deliberately excludes any sector of a community which is entitled to use it. Public libraries make strenuous efforts to attract users from all sectors of the community regardless of economic, social, political or chronological status. Similarly academic libraries try to serve all aspects of the community in their remit as do those in commerce and industry. A concept of deliberate exclusion would be alien to the provision of library and information services generally. Therefore the supermarket model for a library and information service is not valid and, in fact, quite misleading.

The nature of information is also against the adoption of the supermarket model. Although serendipity plays an important part in shopping activities, most people have a fairly good idea of what it is they want to buy or achieve by visiting the supermarket. By its very nature information does not fit into this model. This is because information is actually imparting something to a user which they did not have before and which they did not know and therefore the analogy with shopping is again misleading. Apart from a few basic factual pieces of information most users of library and information services actually use the service to move them on into new avenues of thought or ideas which they have not previously considered.

Considerable time has been spent on the role of the library because it is the primary customer for traditional printed materials. Any changes in the access and availability patterns for information will have a profound effect on the library service and, conversely, any changes in the library services will profoundly affect access and availability.

9.0 PROVISION IN THE ELECTRONIC WORLD

Some time has been spent deliberately on the "normal" publishing models because, until they are fully appreciated it is not easy to understand why and if the electronic world is different. Also it is important to be able to compare the two environments to see if the models can be crossed over from one to the other in any way. Having established in some detail how the current scientific publishing world works in a commercially-driven context we now need to see how these elements are changing in an electronic world. We shall examine each element in the chain to see what relevance it has in the "new" electronic environment.

The role of author, editor, reviewer, publisher, distributor, purchaser and reader are all undergoing major changes, many of which are still in flux and will almost certainly change over quite short periods of time.

Many different models are emerging including the Budapest Open Access Initiative, Open Archives Initiatives and many related models including Open Knowledge Network and Creative Commons. This is a very fast-moving and dynamic area for scholarly communication and any description can be nothing more than a snapshot of the current situation. Those interested can follow some of the more high-powered (and sometimes vitriolic!) discussions on Open Archives, commercial publishing and the role of the scholarly author should refer to Stevan Harnad's discussion list. (This list is called SEPTEMBER98-FORUM@LISTSERVER.SIGMAXI.ORG and was consulted on 18 April, 2002). For example, David Goodman of Princeton University recently wrote:

> Should a rational publisher fear the OAI?
>
> A rational publisher need not and does not fear the OAI for those journals which are worth reading, and consequently worth buying; it will both now and in the future be able to sell these. It certainly ought to fear for those journals that are not worth reading, and consequently not worth buying. The only function of these journals is to certify publications as having been at least superficially "peer-reviewed." This can obviously be done at a much lower cost; thus, if they continue to publish these journals at anywhere near the present prices, they will not be able to sell them. Considering the low demand, it seems probable to me that they might not be able to reduce their costs enough to sell them at all. The solution for a publisher is obvious: it should publish good journals, and only good journals. A publisher complaining about the threat of OAI suggests that it knows very well that the quality of its journals cannot compete. (This comment can be found at SEPTEMBER98-FORUM@LISTSERVER.SIGMAXI.ORG on 1 April 2002.)

However, a proper understanding of the underlying issues raised by these many different and often competing models is crucial to the proper interpretation of the varying ideas being presented. A recent attempt by UNESCO to develop an International Alliance for Information Access (IAIA), focusing particularly but not exclusively on Third-World issues shows how important it is to ensure that these different models, all well-meant and often of extraordinary value and relevance to the distribution of scientific ideas, are co-ordinated in some way to ensure maximum benefit to the communities they seek to serve. At the same time their independence and commercial roots needs to be respected and therefore the most that can probably be achieved is an attempt to provide an overall picture to inform those working in the field rather than to offer any guidance or structure within which these different initiatives can flourish.

Therefore the different roles within the publication chain are themselves in a state of considerable change and the economics of how publication takes place are also under severe strain. How then can scientific information continue to be made available in such an unstable context? Although the answer is that "nobody knows" this is not helpful! So we must try to understand the new roles in the new context as best we can and draw out the inferences for the economics of information distribution accordingly.

9.1 The Author

Whatever else changes in the field of scientific information distribution, there can be no doubt that the author will remain. Without the creativity of the author there can be nothing to distribute. Authors need to make public (carefully avoiding the word "publish" as this has many possible interpretations) their findings and ideas in order to stimulate discussion and move forward scientific and technical development. However, the traditional models described earlier will no longer be viable. There are a number of interrelated reasons for this. These are concerned with personal status, publisher economics, archiving of scientific information, academic and research institutions' attitudes and freedom of access.

To quote the Budapest Open Access Initiative (BOAI) from their website at www.soros.org/openaccess

> An old tradition and a new technology have converged to make possible an unprecedented public good. The old tradition is the willingness of scientists and scholars to publish the fruits of their research in scholarly journals without payment, for the sake of inquiry and knowledge. The new technology is the Internet. The public good they make possible is the world-wide electronic distribution of the peer-reviewed journal literature and completely free and unrestricted access to it by all scientists, scholars, teachers, students, and other curious minds. Removing access barriers to this literature will accelerate research, enrich education, share the learning of the rich with the poor and the poor with the rich, make this literature as useful as it can be, and lay the foundation for uniting humanity in a common intellectual conversation and quest for knowledge.

All of which sounds very good and honourable and idealistic but, as it stands, fails to address the economics of how to make information available.

9.2 Personal Status (Kudos)

This is an essential part of making works available to the public. For this reason the peer-review process is still essential as this guarantees the reliability of the information provided. Publishers of traditional journals agree that it is a costly and time-consuming exercise yet essential to their publications and the integrity of the scientific community as a whole. Failure to incorporate peer-review into journals leaves the field wide open to any eccentric or the deliberate sabotage of serious scientific investigation.

In the present paper-based world the current level of minimal, toll-based access, society is paying an average of $2000 per paper from the minority of institutions that can afford the journal in which that paper happens appear. However, peer review is essential and authors need this to authenticate their work. But peer review has to be in the context of a recognized product, in other words the scholarly journal with its status and reputation. In an electronic context the whole concept of the "journal" has to be maintained in some form in order to achieve this. Merely to put a paper on the web and say it has been reviewed by experts and accepted is insufficient in the present climate of academic approval.

Peer-review is perfectly possible online as well as in traditional methods of distribution. Indeed it is essential it should continue. As Sally Morris of the Association of Learned and Professional Society Publishers (ALPSP) has said:

> People value peer review and they value research being gathered together in things called journals, Peer review will continue exactly as before there was open access. Journals will continue to be journals. What will change is what "gathered in" means. And open access is for those would-be users whose institutions cannot afford access to each given paper, either on-paper or on-line. That corresponds to the majority of potential users, for the majority of the annual 2 million papers in the 20,000 extant peer-reviewed journals. All of that would be lost research impact otherwise (Morris 2002).

9.3 The Economics of Being an Author

Most authors of scientific materials (excluding some textbooks) do not write for money as has already been argued. However, the cost of producing articles for scientific journals is high but not in terms of fees for authors. Authors have more interest in making information available than receiving royalties. To quote the Budapest Open Access Initiative again:

> The literature that should be freely accessible online is that which scholars give to the world without expectation of payment. Primarily, this category encompasses their peer-reviewed journal articles, but it also includes any unreviewed preprints that they might wish to put online for comment or to alert colleagues to important research findings. There are many degrees and kinds of wider and easier access to this literature. By "open access" to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.

In this context authors were not worried about issues such as copyright either as copyright tends to be an economic reward for work done and no such rewards are available in this context. Or are they?

With the introduction of electronic copyright management systems (ERMS) the whole model is changing. The issues of electronic copyright are discussed in another chapter but are also relevant to the economics of making information available. Works issued in paper form could be controlled only by fairly blunt instruments such as licensing agreements which are based on sampling systems to see what is copied and by whom. In return for this the licensee (the person taking out the licence) pays a fee which the licensing agency then distributes to owners of copyright using the data collected by the sampling surveys. When material is distributed in electronic form it is quite possible to determine when an individual article is viewed, downloaded or print out. Consequently exact use statistics can be generated and payment made to the owner on a much more realistic basis. Therefore some scientific authors are now claiming that as publishers make money out of their work the authors themselves should be compensated according to the way their works are used. This was not previously possible.

One consequence of the introduction of electronic distribution is that the role of the licensing agency is weakened. The blunt instrument approach is not needed and the intermediary for collecting fees for copying and use can be eliminated as the technical capabilities of the web expand. Copyright licensing agencies are finding it more and more difficult to retain their position in this context and rely more and more on bulk licensing for educational copying or multiple photocopying in industry with some revenue coming from document delivery services.

The ERMS issue is further brought into play by changing attitudes in many institutions which employ those who write the articles. Formerly they took little interest in ownership of the rights in the material that their researchers produced but there is an increasing awareness of the value of the intellectual content of the material created by their staff. Now, however, such awareness is much more intense and institutions, both in the public educational sector and in the private research sector, are much more keen to obtain rewards for the contribution their staff make to scientific research globally. This value can be assessed and identified using ERMS as this gives a clear indication of the level of use and many researchers will work on the assumption (not necessarily correct) that volume of use = value to the community. The assertion of rights by institutions brings about yet another conflict that between the employer and the employee. To what extent an individual employee owns the rights in what they create is a serious matter of legal dispute in many countries and organizations. It is particularly crucial in the university and higher academic

context where contracts are often unclear (or even silent on the issue) and where the different roles fulfilled by individual researchers are also far from clear-cut.

9.4 Creative Commons

In a boon to the arts and the software industry, Creative Commons will make available flexible, customizable intellectual property licenses that artists, writers, programmers and others can obtain free of charge to legally define what constitutes acceptable uses of their work. The new forms of licenses will provide an alternative to traditional copyrights by establishing a useful middle ground between full copyright control and the unprotected public domain.

The first set of licensing options Creative Commons plans to make available are designed mostly for people looking for some protections as they move their wares into the public domain. Those protections might include requirements that the work not be altered, employed for commercial purposes or used without proper attribution.

Lessig adds that it's possible Creative Commons' licenses may eventually evolve to include options that permit or enable certain commercial transactions. An artist might, for example, agree to give away a work as long as no one is making money on it but include a provision requiring payments on a sliding scale if it's sold. As participation in the Commons project increases, a variety of specific intellectual property license options will evolve in response to user needs, which in turn would create templates for others with similar requirements (Lessig 2002).

9.5 The Publisher in the Electronic Context

It must be acknowledged that publishers are nervous at the present time. There is a great swell of opinion that the traditional publisher and publishing methods are doomed because of the advent of the web and instant access to individual works by individual researchers. Whilst nobody would argue with the fact that publishers publish for profit, the basis for the activities of the industry is to make the creativity of authors more widely known to the public. Indeed, away from the recreational writing market, most authors write not for profit but to make ideas more widely known. But no author or publisher can reach all potential readers of any work and needs intermediaries. Although booksellers fulfill part of this role they provide only those materials which are likely to sell and therefore their role as intermediaries is limited to commercially attractive material. The library provides the interface between the publisher and the untapped, and untappable market and therefore enables one aim of publishing – to reach the public – to be more effectively achieved. Publishers cannot hope to reach every potential outlet for their products because they do not have the direct contact with the necessary groups to achieve this. Publishers are also limited in what they can provide in terms of a repertoire which will normally be limited to their own products or those of associated companies.

The argument is that all this changes in an electronic world. Publishers can reach anyone who has the technological capability to connect to their database and online services. Publishers should view this new model with excitement as it opens up a huge potential market never before realized.

However, publishers also realize that the majority of their current products are not bought or paid for by individuals but by institutions, especially libraries. If publishers are to market their products direct to the end-user then a revolution in economic modeling needs to take place.

9.6 Pay Per View and Pay As You Go

Electronic methods of providing access enable users to choose more directly exactly what it is they wish to view or use and not be required to scan through material that is neither relevant nor useful to them. It also

enables subscribers to pay only for what they actually want rather than buy paper journals which, as predetermined packages, deliver what the publisher wishes to deliver and which often contain unwanted information of many kinds. Given that many articles are never red at all (an average readership of 0.2 per journal article was found by an ICSTI study (ICSTI 1996) some years ago) this is plainly a more satisfactory way of paying for information. However, it does mean that the actual costs of access are infinitely variable and not easy to predict, especially if publishers vary the cost of access and use within a given journal title. This is entirely possible given the sophisticated technology for controlling access now available. In the paper world, once a journal has been bought, the level of usage is important only in determining if it is value for money, otherwise limitless use is possible, albeit on a one-person-at-a-time basis. The electronic methods of publishing allow for concurrent multiple access Moreover, access can be site-wide or even on multiple sites simultaneously provided the use is allowed and paid for. This scenario raises a number of important questions.

Firstly, payment for access. The vital question is: who is going to pay for access to electronic materials if they are to be individually accessed by researchers and even students direct. Although those in a purely commercial context will probably be able to afford access to whatever they need, because of the economics of information use in industry, others may well be disadvantaged because of the complex pricing models that copyright owners impose through the electronic world. High access charges may be one difficulty for libraries anyway but increasingly copyright owners are demanding payment before use is permitted. Whilst the library may be able to manage this to some extent by “up-front” payments, very often the use will be determined by the user on an ad hoc basis.

A public or academic library may provide online access free of charge to all its users but how does the library manage the situation where, for example, a user begins a search on the Internet, only to be asked for a credit card number before they can proceed to a particular database. The librarian has to face the challenge: do we continue to struggle to provide access to information which will be free of charge at the point of delivery to the user or do we have to develop alternative service models which either limit the access any user can have or require the user to pay beyond a predetermined limit. The problems of internet publishing have been explored extensively without any major conclusions being reached (Kahin and Varian 2000).

Library budgets throughout the world are not growing. How many readers of this paper would be able to say that their library budget is actually growing? Not many people would answer this question with the word “yes”. Perhaps we have to look at alternative models from other parts of society. If a student in a university cannot pay their fees, there may be financial support from the local or national government, from charitable funds or from the university itself. In public libraries we may need to look at how citizens are funded to obtain satisfactory housing, free public transport, free medical care or reduced charges for using local sports or leisure facilities. In a world where smart cities are rapidly becoming a reality, it may be time for the services offered by library to be included in this scenario. After all, in some cities one smartcard will allow you to pay your local taxes, provide access to local transport (at full rate or reduced/free tariffs if the user has special social needs), give entry to the swimming baths or even pay your library charges! Perhaps access to information needs to be seen in the same context as other social services which are paid for by the local, state or national government. If you entitled to subsidized housing, perhaps you should also be given subsidized access to information.

10.0 THE ROLE OF THE LIBRARY

On the other hand, libraries are used by millions of people who never had any intention of buying their information nor have the economic power to do so.

Libraries, however, can, and do, reach a wide audience as they have direct access to a very broadly based user community. They can also offer a much wider range of products than the publisher or even other

intermediaries (booksellers, subscription agents, database hosts) as they are not motivated primarily by financial incentives, although they may need to limit the range of resources available because of financial constraints. In the case of public libraries and the majority of academic and educational libraries, this achievement is at the expense largely of the tax-payer throughout the world.

Naturally libraries, as major resources of information, and valuing their unique role to reach so many users, want and need to be able to exploit new possibilities. However, if they are no longer going to rely on a paper-based industry, the alternative will be to use materials in electronic formats of many kinds, all of which are vulnerable to a range of threats including unauthorized copying, redistribution, repackaging and even republishing under different labels. Document suppliers, whether in traditional libraries or in the commercial sector, will certainly need to be able to do some of these things, to meet the changing expectations of their customers.

The changing role of the library in the economic model for publishing is explored further in the chapter on intellectual property rights.

11.0 ACCESS VERSUS OWNERSHIP

In the paper world subscribers to scientific journals purchase a physical volume and can do as they like with it (technically called exhaustion of rights). In the electronic model the economics are quite different. There is nothing to "own" in the usual sense. The publisher owns the electronic database in which journal articles are stored and allows access to this either on a pay-per-use basis (in which case the user obtains possession of either a printout of the article or simply the right to look at it onscreen for a limited period) or access on a predetermined level from a given number of access points for a specific geographic area and for a specified length of time in return for licence fees which may be fixed in advance in accordance with the size of the elements just mentioned. Licences can be highly complex but will usually cover such issues as

a) Number of access points
b) Number of parallel uses
c) Actions allowed – downloading, printing, storing, networking (intranet) re-use
d) Archival access
e) Time before licence is renewed or re-negotiated
f) Access in the event of the licence being terminated

In traditional publishing a) and b) are determined by the number of copies bought; c) is limited to photocopying and may be negotiated or, more likely, carried out under national copyright laws; d) is not an issue as the purchaser has possession of the physical publication; e) is simply a matter of renewing a subscription or not and f) is like d). In an electronic model all of these issues are essential. a) and b) will help determine the price of the licence; c) will be vital as the list of actions (which is not exhaustive) will govern the price but also be important for the subscriber to be able to carry out more sophisticated information provision processes than is possible with paper publishing), d) is a matter of major concern as access to the **total** electronic file is governed by the fact that the institution subscribes at all. Rarely are institutions allowed retrospective access to databases to which they no longer subscribe yet archival material is essential to good scientific research and investigation. e) allows a certain element of stability in providing access and f) is an extension of the archival problem.

12.0 LICENSING AGENCIES

One way in which publishers try to gain some benefit from copying in traditional media is the licensing agency. Agencies for collecting royalties go back a long way. In the UK the Performing Rights Society,

which collects royalties for composers whose works are performed live in public, was founded in the 1920s. Many countries now have collecting societies for published books and journal articles. Their methods of operation vary but, generally speaking, they offer a licence to users to copy material in return for fees which may be transaction-linked (i.e., each individual action is recorded and used as the method for distributing the royalties collected) or they may offer a "blanket" licence and use sampling methods to determine how the royalties should be distributed. The former is cost-intensive and requires detailed record keeping but pays owners what they should get; the latter is much cheaper to administer but offers a sort of "rough justice" so that owners get something which approximates to their rightful payments but not exactly and this will vary from one sampling point to another. Agencies exist to cover published books and journals, performances, newspapers and in some countries film rights as well. In the "physical" delivery of information the licensing agency has a vital role to play but in the electronic context owners do not see the benefit of this "middle-man" in the payment chain. As it is now possible to monitor use of electronic material exactly and even arrange online billing and collection of royalties publishers see the licensing agency as irrelevant. Their role in electronic royalty collection has been limited largely to licensing the digitization of existing paper materials rather than being involved in "born digital" materials. The great disadvantage from the user's point of view is that they are required to have separate agreements with many different suppliers whereas licensing agencies make this unnecessary in the same way that periodical subscription agents reduce the number of individual transactions that libraries need to perform. Whether or not the licensing agencies will be able to transform themselves into Trusted Third Parties (TTPs) which collect money on behalf of a range of owners from many different users is still open to question.

12.1 International Coalition of Library Consortia

The ICOLC have issued various guidelines to try to establish "ground rules" for publishers and libraries in acquiring and making available electronic journals under licence. They say, for example: that publishers price most e-journal content using the cost of the print publications as their base price (the "print-plus" model).but in many offers from publishers, the pricing of the electronic journal is expressed as an "add-on" to the price of the print product, or alternatively the price quoted is linked to a "no-print cancellation" clause in the contract. A few publishers now offer an "electronic-plus" model, with the electronic journal being supplied for a base price and a price for print copies being added to that base price. This model is acceptable, provided that the purchase of the print copies is optional, ***and*** the base price for the electronic content is no more than a reasonable percentage (say 80%) of the price for the electronic-plus-print (thereby reflecting the savings that the non-supply of print copies can bring), ***and*** the combined electronic and print price is no more than current print-only prices (thereby reducing the risk of additional cancellations to pay for both formats).

ICOLC also emphasizes that their members wish to receive from publishers offers that are not necessarily based upon the traditional title-by-title subscription model. Some publishers have already broken away from that model and offer their total content for the price that a library might have paid for a limited number of print journal subscriptions. This "all-you-can-eat" model does meet the needs of many – although not all - libraries and consortia. However, to meet the diverse needs of different consortia and libraries, "all-you-can-eat" should be one of the options offered by a publisher. The wider the number of choices, the greater the chance of satisfying the customer. It is hoped that publishers will be encouraged to offer additional pricing options that provide increased value for money in certain situations, such as "pay-as-you-use" options by which the consortium or library may purchase blocks of journal articles, or may pay only for delivery of the articles that are actually used, or "Total title purchase" for selected groups or subject clusters of titles, with "pay-as-you-use" available for the titles not selected.

One major concern to libraries and institutions generally is the "no-print cancellation" clauses in licenses and contracts for e-journals, and to pricing models that impose financial limitations or penalties when

cancellations are permitted. It is an increasing view that publishers should direct more effort toward new pricing models that break away from print-based models, as explained in this document.

Publishers see the electronic world as one in which they can, for the first time in many cases, actually determine what is used and by whom. This is valuable marketing information but it raises two important issues.

Firstly, confidentiality. Use by individuals of scientific material could reveal all kinds of personal data apart from straightforward marketing information. Reading patterns could be used to reveal political or philosophical interests, sexual orientation or health data. Secondly, the use of scientific literature by a research institution or company will give clues as to its commercial programme and exploitation.

Most subscribers to scientific journals are interested in working with publishers to develop purchasing models that meet the diverse needs of educational organizations in different countries. For example, the traditional print-based subscription model may become less satisfactory for meeting the educational needs of North American and Western European institutions. Innovative partnerships between educational authorities and publishers may be needed to produce a better service to users in countries in all parts of the world.

It is also encouraging to note recent initiatives from some publishers who provide electronic journals for free or at very affordable access to countries in transition, such as the programs to provide health-related information through the World Health Organization, and the science and technology publisher responses to the Open Society Institute Electronic Information for Libraries (eIFL) program. Publishers are encouraged to also address the needs of more developed nations that may be experiencing extremely weak national currencies. These issues are discussed in more detail on the ICOLC website at http://www.library.yale.edu/consortia/.

13.0 THIRD WORLD SITUATION

If the problem of "who pays" is of concern to those working in relatively developed economies, how much more worrying is it for those in economic situations which are seriously under-developed or still trying to develop to higher standards. The issues raised are enormous for such countries. Capital is limited and budgets for research and scientific development are easily reduced in the light of any economic changes. In addition, there are often basic issues of currency exchange and lack of availability of "hard" currencies with which to pay for access and the more mundane but crucial problems of whether the technology is available to access the material in the first place and, if it is, does it work. A number of initiatives in this area are trying to find economic models which will allow Third World countries to have access at much reduced costs. Mention has already been made of IAIA and its founding document states

> Information is a basis for knowledge, and equitable access to information needed for human development is a basic human right in all branches of education, science and culture. All members of the Alliance commit themselves to enabling access to essential information that is appropriate, affordable, of guaranteed validity and quality, and in formats that respond to the needs of the intended audiences. Such information and knowledge can originate from any country of the world and be in any language.
>
> Open Access initiatives have arisen in a number of sectors. These have the common objective of working to allow users, particularly those in developing countries, to obtain information products that would otherwise be unaffordable, and therefore inaccessible. These initiatives operate in an information environment that includes 1) sources of information (publishers, authors and other rightsholders), 2) information transfer media (the whole gamut of online and offline options), 3) rules for information transfer (technical, economic, legal and

> normative rules), and 4) users of information (whether infomediaries such as libraries, documentation centres, and teaching establishments, or individuals), who may themselves also be sources of information. Open Access initiatives typically include provision for a two-way flow of information.

Such initiatives include INASP (International Network for the Availability of Scientific Publications), The Essential Electronic Agricultural Library (TEEAL) and International Programs of American Association for the Advancement of Science. More information on TEEAL can be found on their website at http://teeal.cornell.edu.

Some publishers are now aware that the economics of information supply to developing countries cannot meet the needs of those countries nor can they provide the publisher with a sustainable return on investment. Initiatives by major scientific publishers in the health field, such as Academic and Elsevier, have resulted in much reduced subscriptions for developing countries. This programme has been organised in conjunction with the World Health Organization (WHO).

14.0 SPECIFIC EXAMPLES OF LICENCES AND MODELS

Emerald Press (formerly MCB University Press) have taken the initiative for a number of years by providing subscriptions which, although high by most publisher standards, include all kinds of additional benefits to the purchaser. In Emerald's own words:

> The pressure on libraries to provide more information with less budget is growing. Academic librarian professionals need to find ways to satisfy the research needs of increasing numbers of faculty and students; while managing new technology, increasing workloads and decreasing staff numbers. Library consortia networks have existed for a long time, particularly in the USA, but their main aim of sharing printed material, has changed to providing common access to electronic resources via the Internet. Emerald has developed a consortia model that provides a win-win solution for both publisher and library."

Emerald Consortia initiatives provide:

- Lower cost information
- Access to considerably more journals and information
- Price predictability and budget stability for fixed terms
- Purchasing options to suit everyone
- Preferential pricing and discount for additional paper and electronic resources
- Flexible licensing models
- Library resource facilities and training
- Archiving solutions.

By buying into the Emerald package any subscriber can obtain all or any of these benefits. Naturally the pricing reflects the range of possibilities but it goes a long way to achieving what has been said earlier in this paper – that libraries and institutions need flexibility to be able to deliver services the publisher cannot or is unwilling to do because of the economics of the service.

JSTOR is a not-for-profit organization in the U.S. dedicated to long-term preservation of and access to scholarly publications. And more information about its activities can be found at http://uk.jstor.org/about/need.html JSTOR Arts & Sciences I Collection includes non-current issues of 117 important research

journals. Journals are digitized back to the first issue published (many of which date from the 19th century) and continue to a date no more recent than 3-5 years prior to the most current published issue. The licence sets out both permitted uses and prohibitions and is quite generous in its scope. Fees for the licence are determined on a banded scale according to the size and nature of the institution. For example the licence sets out Permitted Uses which puts into context what JSTOR's objectives are. The collections are provided for educational purposes only. Publication, redistribution to persons other than Authorized Users and all commercial use is expressly prohibited. Products, unless there is a clear statement to the contrary, are licensed for use only by faculty, staff, students, and "walk-in "users of the licensed institution. The agreement permits use by specific categories of users only. In the case of JSTOR these are detailed as

a) Students, Faculty and Staff regardless of location
b) Remote Access within the country concerned
c) Remote Access Overseas
d) Student Placement in Industry
e) Walk-in-Users
f) Honorary Members of Staff.

At the same time the agreement prohibits use of various categories as well, for example:

a) People from Industry on Courses Run by their Employer
b) Retired Members of Staff
c) Alumni.

Having set out WHO can use the service the licence then details what they may or may not do. Authorized Users may, within the scope of the JSTOR User Rules:

a) Access and use JSTOR Collections for classroom instruction and related activities including handouts, presentations, research, and student assignments.
b) Use JSTOR Collections as part of a professional presentation at a conference, seminar, workshop, or other professional activity or in a public display or performance in the (Institution name) gallery or similar facility.
c) Use JSTOR Collections for student or faculty portfolios, term papers, theses, and dissertations, provided copies of these are not published or redistributed.

The licence further indicates that authorized Users may not:

a) Use JSTOR Collections for any purposes other than education, research, or scholarship.
b) Use JSTOR Collections for any commercial or business-related purpose whatsoever.
c) Reproduce, distribute, redistribute, or publish JSTOR Collections outside of your institution without obtaining permission.
d) Download from the JSTOR archive an entire issue of a journal, significant portions of the entire run of a journal, or a significant number of sequential articles unless prior written permission has been obtained from JSTOR.

These are detailed but quite reasonable and enable the user to know whether or not they are allowed to use the system and, if so, what they are permitted to do. In return for this JSTOR charges a fee and subsequent individual uses of the material are of no interest provided they are within the rules.

15.0 ALTERNATIVE MODELS

Many attempts are being made to produce alternative models and alternative publications. For example the services offered by BioMed Central. As it states on their website www.biomedcentral.com which says that I/we retain copyright.

I/we grant to any third party, in advance and in perpetuity, the right to use, reproduce or disseminate the article, in any format or medium, in whole or in part, provided that the integrity of the article is guaranteed and not compromised in any way, that BioMed Central is duly identified as the original publisher, and that proper attribution of authorship and correct citation details are endorsed on the article or its parts.

I/we grant to BioMed Central (its successor and assigns) an irrevocable world-wide licence for the full term of copyright in the article to publish it and identify itself as the original publisher.

This respects the rights of the original creator but, at the same time, releases the material to be used by BioMed. Those rather "ethereal" moral rights (paternity and integrity) become firmly fixed in a contract for publishing in an electronic context. This increasing emphasis on moral rights is crucial in this search for alternative models as the search so often focuses on the needs of the individual not the institution or the publisher. Therefore to focus on individuals for their needs but not their rights would be bizarre.

16.0 SCHOLARLY PUBLISHING MODELS

SPARC (the Scholarly Publishing and Academic Resources Coalition) has launched *Gaining Independence: A Manual for Planning the Launch of a Nonprofit Electronic Publishing Venture*. This is a detailed, step-by-step guide leading readers through the creation of a business plan for start-up and early-stage electronic publishing ventures, including digital repositories and journals. Gaining Independence will help universities, libraries, societies and others conceive, plan and implement alternatives to commercially published scholarly and scientific information. It provides background on relevant electronic publishing models and focuses especially on areas of business planning that may be unfamiliar to those considering new communication initiatives. The manual includes sections on: Situational Assessment and Strategic Response; Technology and Technical Considerations; Markets, Marketing and Sales; Organization; Finances; and the Financial Plan and Operating Plan.

SPARC was founded as a constructive response to market inequities in the scholarly communication system and is taking steps toward building a system that serves the needs of the scholarly community and facilitates effective partnerships between scholars and their institutions or societies. Our aim for Gaining Independence is to help make alternative scholarly initiatives mainstream and self-sustaining by emphasizing the application of sound business planning practices.

17.0 CONCLUSION

The economics of publishing are in a state of flux. Publishers, authors, libraries and users are all unclear about their new roles in a business which is being changed out of all recognition by the advent of the worldwide web and other electronic methods of communication. Authors need to decide what they wish to achieve by being published, publishers must consider how, and if, they can generate profit in this new context; libraries will remain but will not remain the same (Smail 2002). Those funding research and innovation must consider whether they wish to generate income from the innovation itself or also from publishing information about it. Questions abound: will traditional publishing survive? Can commercial publishing continue in science at all? Will online access totally replace the printed word? Who will pay for access in the future? Who will pay for those unable to afford access? This last question has far-reaching implications not just for the Third World but for those strata of society in developed economies where

individuals need information but cannot afford to pay for access to it. Nobody knows the answers to these questions and only time will tell what they will be. Everyone involved in information creation, provision or use needs to be aware that they are waiting to be answered.

18.0 REFERENCES

Harvey-Jones, J. (2002, March 20). Why the future looks bleak for books. The Independent, p. 4.

ICSTI. (1996). (International Council for Scientific and Technical Information. A comparative study of access to journals through subscriptions and document delivery. Paris, ICSTI.

Kahin, B. and Varian, H.R. (2000) Internet publishing and beyond. the economics of digital information and intellectual property. Cambridge, Mass.: Harvard University.

Lessig, L. (2002). Custom licensing provides much-needed middle ground for content. Eweek, February 13, 2002.

Morris, S. (2002). Interview with BBC Radio. April, 2002.

Smail, C. (2002). Providing access through co-operation: summary of a conference. Interlending & Document Supply, 30(1): 32-34.

Metadata for Electronic Information Resources

Gail Hodge
Information International Associates, Inc.
312 Walnut Place
Havertown, Pennsylvania 19083
USA

gailhodge@aol.com

ABSTRACT

The rationale for cataloging and indexing of electronic information is much the same as for print materials. Cataloging and indexing provide a surrogate for the item, which facilitates resource discovery and access. But, what has changed in the electronic information environment is the terminology. In the Internet environment, the terms cataloging and indexing have been replaced with the term "metadata." Metadata is often defined as "data about data" or "information about information." The term, which originated with the data and computer science communities, is now in general use for the cataloging and indexing of electronic information sources.

Metadata serves three general purposes. It supports resource discovery and locates the actual digital resource by inclusion of a digital identifier. As the number of electronic resources grows, metadata is used to create aggregate sites, bringing similar resources together and distinguishing dissimilar resources.

There are a variety of metadata schemes that serve different purposes for different object types, subjects and audiences, including the Dublin Core, Metadata Object Description Schema (MODS), the Global Information Locator Service, the Text Encoding Initiative Header, the Encoded Archival Description, the Content Standard for Digital Geospatial Metadata, the Data Documentation Initiative, and the draft Technical Standard for Still Images. A metadata scheme has three components – semantics, content and syntax. An extension adds elements to an existing scheme to describe a particular resource type, handle material on a particular subject, or address the needs of a particular user community. Profiles are subsets of a larger scheme that are implemented by a particular user community. Metadata can be embedded in an electronic resource or stored in a separate file.

A growing number of tools, both open source and commercial, are available to create and edit metadata. Creation may be done manually or by metadata generators that extract key information from the object. Metadata harvesters capture metadata records that have already been created using the "shared cataloging" model. While many projects aimed at having metadata created by the object's author, this has proved to be difficult to implement. An alternative is to have a core set of metadata created by the author with editing and quality control performed by a librarian or editor who has a view of the whole collection.

With disparate metadata schemes, ensuring that information collected in a specific scheme by one organization for a particular purpose can be exchanged, transferred or used by another organization for a different purpose becomes an issue. Metadata frameworks, crosswalks, and registries are ways to achieve interoperability.

Use of controlled and uncontrolled vocabulary terms is encouraged, particularly within specific subject domains. However, most metadata schemes do not dictate the use of a particular controlled vocabulary but instead allow the vocabulary scheme to be defined within the syntax.

In order to increase the use of metadata, systems that support metadata creation and search engines that take better advantage of metadata must be developed. Communities of practice should develop content standards, along with other groups that share common interests. Stakeholder groups must be made aware of the importance of metadata for the short and long-term enhancement of the electronic environment.

1.0 THE PURPOSE OF METADATA

Similar to traditional cataloging and indexing, metadata performs three main functions. It facilitates discovery of relevant information, locates the specific resource, and organizes electronic resources into collections. In addition it provides information needed to administer and manage the collection. Technical metadata is needed to allow digital objects to be re-presented in new technical environments.

1.1 Resource Discovery

One of metadata's primary functions is to support resource discovery by describing aspects of the original electronic resource in which the designated user community may be interested. Metadata, such as titles, subject terms and abstracts or descriptions, are particularly important for electronic resources, such as datasets or photographs, that have little if any text content on which current text-based Web searching can be performed.

Metadata can describe the resource at any level of aggregation – a single resource; a part of a larger resource, for example, a photograph in an article; or a collection of resources, such as a digital library. The level at which metadata is applied depends on the type of data and the anticipated access needs. Datasets are generally cataloged at the file or collection level. Electronic journal articles may be cataloged individually, sometimes with no concern for metadata at the issue or journal title levels. Generally, the metadata for Web sites is applied to one or more pages that make up a cohesive resource with informational value.

1.2 Location of Electronic Resources

Once a resource has been discovered via the metadata, the resource must be located. Metadata supports the location of the actual digital resource on the network. Most metadata schemes include an element that is defined as the unique identifier needed to locate the resource.

In practice, most metadata schemes continue to use the URL, or the Uniform Resource Locator, as the unique identifier. The URL is the physical address, the server or domain name, directory and file name for the resource. This provides fast look-up but is problematic as the Web grows and information managers need to move the physical locations of the resources. In the case of electronic journals, URL changes may occur due to the merger or acquisition of one publisher by another. URLs that are not up-to-date or that have not been forwarded to a new URL result in the famous 404 message indicating that the Web page cannot be found.

In an effort to solve this problem, two major systems have been developed. First, OCLC developed the Persistent URL. This method continues to use the URL construct, but it sets up a resolver service. The PURL is used in the metadata record or in reference links that refer to the electronic resource. When a browser attempts to locate the PURL, it accesses the record in the PURL Resolver service at OCLC. The Resolver uses standard Internet redirection to access the actual URL of the resource's physical file location. If the location for the actual page changes, its owner must change the URL in the Resolver, but the PURL that has been published remains the same.

The PURL is structured as:

http://purl.oclc.org/[specific resource file name]

The beginning of the PURL is the URL for the PURL Resolver Service (in the example above, the resolver at OCLC is used) and the file name in brackets is the file name for the specific resource.

The second method is the Handle System® developed by the Corporation for National Research Initiatives (CNRI) under contract to several U.S. government agencies. In the Handle, the prefix is a unique identifier assigned to the resource owner by the central Handle System. This prefix ensures that the identifier is unique. Following the slash is the suffix assigned to the item by the producer.

A Handle is structured as:

[unique prefix for the assigning agency]/[persistent, unique identifier for the resource]

The unique identifier in a Handle can be any item ID. Possibilities include the ISBN, the Standard Item Contribution Identifier (SICI), the Publisher Item Identifier (PII), or a local accession number.

The Handle also uses a resolver service, but it allows more flexibility in the structure and syntax of the identifier. Because it actually uses a database scheme, a single Handle can resolve to multiple locations for different versions of the same resource. Different versions of an electronic resource, for example one in HTML and another in pdf, can be uniquely identified even though they have the same Handle, because the database also contains the data type. The data types can be resolved based on a user's preference or an interface can be designed that offers the user a choice between the versions.

The Handle is the underlying technology for the Digital Object Identifier. The DOI, managed by the International DOI Foundation, establishes a specific syntax for the DOI under the Handle framework. The DOI is the basis for a system called CrossRef. CrossRef is a DOI Registration Agency formed by a consortium of electronic journal publishers. The members of CrossRef deposit their DOIs into a central repository maintained by CrossRef. The purpose of CrossRef is to facilitate linking between electronic journals, primarily from the references at the end of an article to the full text for those articles. The DOI in CrossRef is used to form the reference link from a reference to the full text article. The CrossRef service is particularly valuable when the references are to articles from a publisher other than that of the referring journal.

As mentioned earlier, CrossRef is a DOI Registration Agency, which maintains a central repository of DOIs in order to allow publishers to move their physical files, while maintaining a persistent link in previously published references. In addition to the DOI itself, CrossRef maintains a minimal set of metadata for each DOI. This limited metadata, consisting of the article title, the first author's last name, and journal citation information, allows a publisher or library to find the DOI for an article published by a member of the CrossRef system in order to embed the DOI in a reference or to implement linking services across resources which the library has licensed.

1.3 Organization of Resources into Collections

In addition to the discovery of specific resources, metadata brings similar resources together and distinguishes dissimilar resources. As the number of Web-based resources grows exponentially, aggregate sites, portals, or subject gateways are increasingly useful in organizing resources based on audience or topic. Originally,

these resources were built as static Web pages with the names and locations of the resources "hard coded" in the HTML. However, it is more efficient and increasingly more common to build these pages dynamically from metadata stored in databases.

Content management systems support the development of such portals by managing individual digital objects. Metadata created as part of the content creation process is used to select and organize individual digital objects into different portals by subject, business function (accounting versus manufacturing), or other organizing principle. Metadata information is also matched against user profiles to create customized (MyLibrary or MyPortal) Web sites.

Another method of organizing Web information is through channels. Channels are pre-selected Web sites that automatically "push" collections of information to a user's browser. They are commonly used for continuously updated information such as stock quotes and news. The dominant metadata scheme for webcasting is the Channel Definition Format (CDF) developed by Microsoft and its partners. The CDF provides metadata elements such as the title of the channel, an abstract, the publication date, the last date the content was modified, the logo for the channel and the schedule on which the channel's content is updated so the "pushing" can be scheduled.

1.4 Administration of the Collection

A fourth type of metadata is used to manage the digital object and its metadata. The elements that may be found here depend in part on the workflow for the creation, capture and long-term use of the digital object that is being archived and preserved. They include control elements such as the date created, the date captured, the operator, and the date last migrated.

1.5 Presenting Content in New Technical Environments

Technical metadata is the overall term used for metadata elements that describe the computer hardware and software needed to reproduce the digital object. This includes file formats such as pdf and video formats such as mpeg. These are connected to the readers or browsers that must be available for a user to be able to access the object. This set of elements is often considered part of the preservation metadata set because it is critical to rendering the digital object in new technical environments in the future or when using emulators of obsolete technologies. Technical metadata schemes are often quite large and detailed, since they are often intended for use by technicians or for computer to computer communication.

1.6 Other Metadata Functions

There are other applications for metadata that cut across the functions described above and often result in specific element sets. Digital rights elements indicate who owns the object and what rights various groups have to use or reuse that object. Rights elements may also include security classifications or distribution limitations. There are several schemes that have been developed particularly in the music and learning objects communities. In systems, the rights management elements must be matched against profile of the user (following proper authentication) in order to ensure that the material is being properly distributed and in some cases the proper payments are being made to the rights holders. The variety of systems, the potential economic impacts, and the variety of materials requiring rights management have led to the concept of a Digital Rights Expression Language (DREL) that is of broad applicability and that can be used by a variety of automated systems in e-commerce. IEEE, MPEG21 and others have been working on rights elements and expression languages.

Preservation metadata or metadata to support the provenance of an object and the preservation of the object is another cross-cutting function. Some of this information is also handled by technical metadata, metadata for discovery, rights management, etc. The current work in this area is discussed in the lecture paper on Preservation and Permanent Access.

2.0 BASIC METADATA STRUCTURE

This section describes the general structure of a metadata scheme, the modification of a scheme to increase its flexibility and usefulness by various communities of practice, and the storage of metadata.

2.1 Components of a Metadata Scheme

A metadata scheme (also called schema) is made up of three structural components – semantics, content and syntax. The definition or meaning of the elements is known as the semantics, and includes the tag set for the elements. For example, a scheme for a text resource may define the Title element with a tag of TI. Generally, the semantics of a metadata scheme are grouped into three types – descriptive, structural, and administrative. These types are complementary to the basic reasons for metadata described above. Descriptive metadata identifies a resource for purposes of discovery and identification. It includes elements such as title, abstract, author, and keywords. Administrative metadata provides information to help manage a resource, such as when and how it was created, its file type and other technical information. Structural metadata indicates how compound objects are put together or how this resource relates to others in the collection.

The set of preservation metadata currently being developed by the Research Libraries Group and OCLC includes elements from all three of these semantic types, but it adds elements specific to preservation activities such as the provenance of the item, the preservation strategies employed, its migration history, etc. (Preservation metadata is discussed in more detail in the session on "Preservation of and Permanent Access to Electronic Information Resources.") Technical metadata is generally considered to be a subset of preservation metadata, because it provides information needed to successfully manage an object through a variety of technological changes and to render the object in new environments.

The scheme may also specify syntax rules for how the elements and their content should be encoded. Metadata can be encoded in MARC21, in "keyword=value" pairs, or in any other definable syntax. Many current metadata schemes use XML (Extensible Mark-up Language). A metadata scheme with no prescribed encoding syntax is called "syntax independent."

The third structural component of a metadata scheme is the content, or the values used to complete the elements. A scheme may specify rules, also called a "content standard," for the formulation of the content (for example, how to identify the title) or rules for the representation of the content (for example, capitalization, language or transliteration rules).

2.2 Extensions and Profiles

Specific implementations or the needs of a certain community can result in modifications to a metadata scheme. Since it is often difficult to anticipate the ways in which a scheme might be used, schemes that can easily be modified are preferred over those that are more restrictive. Modifications are of two types: extensions and profiles.

An extension is the addition of elements to an already developed scheme to support the description of a particular resource type, to handle material on a particular subject, or to address the needs of a particular user

community. Profiles are subsets of a larger scheme that are implemented by a particular user community. Extensions generally increase the number of elements that can be used; profiles constrain the number of elements, refine or narrow the definitions of certain elements, or specify the rules for completing the content of certain elements.

In practice, many applications use both extensions and profiles of base metadata schemes. The metadata scheme for the U.S. Department of Education's Gateway to Educational Materials (GEM) Project is based on the Dublin Core. However, GEM limits the elements to be used (for example, Contributor is not used). It also extends the Dublin Core element set by adding elements that are important to the educational community when describing and using educational resources. These fields include audience (teacher versus student), grade level, and relevant educational standards.

Similarly, the Visual Resources Association (VRA) has established core categories (or elements) to describe visual materials such as buildings, photographs, paintings and sculptures in visual resource collections of slides or photographs. Therefore, metadata for these materials must accommodate the description of the same resource in different media, for example, the original painting, a slide of the painting, and a digitized image of the slide. The VRA Core Category scheme, a profile and extension of the Dublin Core, consists of 17 optional metadata elements: record type, type, title, measurements, material, technique, creator, date, location, ID number, style/period, culture, subject, relation, description, source, and rights. The Dublin Core Relation field is used to relate the records for the same resource in different media. The VRA Core scheme does not specify any particular syntax or rules for representing content. Managers of visual resource collections hope that use of the VRA Core Categories will allow them to share descriptions of original works as well as to better describe materials in their own collections.

2.3 Metadata Storage

Metadata can be embedded in an electronic resource or stored separately. For example, metadata is often embedded in HTML documents as metatags or in the headers of image files. The use of HTML metatags specifically may make the content of the metadata accessible to Web search engines. Storing metadata with the resource ensures the metadata will not be lost, eliminates problems of broken links between the resource and its metadata, and facilitates updating of the metadata and the resource.

However, sometimes it is difficult to embed metadata in certain types of resources. In these cases, storing metadata separate from its electronic resource simplifies the management of the metadata. External metadata may also facilitate search, retrieval and exchange of metadata with other systems and organizations. External metadata is stored in a Web-accessible database system (often called a clearinghouse or catalog) and then linked to the electronic information it describes by a URL or other identifier in the metadata. Implementations that take advantage of the hierarchical nature of RDF and the expressiveness of XML, as opposed to more structured database technologies, may "wrap" the digital object with the metadata to more closely marry the metadata record with the actual object.

3.0 METADATA SCHEMES

Metadata schemes (also called "schema") have been developed and defined by a variety of communities, for different purposes, and for different types of electronic resources. This section describes some common metadata schemes. In addition, some lesser known schemes have been selected to show the range of electronic resources and purposes for which schemes have been developed. While the focus here is on electronic library

resources, it should be noted that many other metadata schemes have been developed in support of e-commerce and electronic data exchange.

3.1 Dublin Core

The Dublin Core is perhaps the most well known metadata element set. The original objective of the Dublin Core was to define a set of elements that could be used by authors to describe their own Web resources. A few relevant elements and simple rules were defined so that non-catalogers could provide basic information for resource discovery.

Dublin Core 1.0 consists of 15 elements: title, subject, description, source, language, relation, coverage, creator, publisher, contributor, rights, date, type, format, and identifier. Recently, the audience element was defined to support the broad needs of the educational and learning object communities. All Dublin Core elements are optional and all are repeatable. The elements may be presented in any order. Note that in the following example relation, contributor and source are not applicable and so they do not appear.

Dublin Core Elements For This Paper

Title: Metadata for Electronic Information Resources

Creator: Hodge, Gail

Subject: metadata

Description: Describes metadata standards and projects.

Publisher: NATO

Date: 20040601

Type: Text.Report

Format: text/html

Identifier: http://www…….

Language: English

Coverage.Spatial: International

Rights: Copyright 2004, Gail Hodge

Audience: Technical

While the Dublin Core description recommends the use of controlled values for fields where they are appropriate (for example, controlled terms from a thesaurus for the Subject field or the use of the ISO language names and abbreviations for the Language field), this is not required. The content rules are determined by the particular implementation, but the adoption of profiles that define domain-specific rules is encouraged.

The Dublin Core was developed to provide simple and concise descriptions specifically to support the resource discovery of Web-based documents. However, in part because of its simplicity, the Dublin Core has been used with other types of materials and for applications demanding increased complexity. The desire to be able to specify more detail resulted in unqualified (or simple) Dublin Core versus qualified Dublin Core. In qualified Dublin Core, qualifiers are used to refine the meaning of an element or to specify the domain

values or rules for representing an element. The element "Date", for example, can be used with the qualifier "created" to narrow the meaning of the element to the date the resource was created. A qualifier can also be used in the element "Date" to specify the ISO 8601 standard as the required format for representing date.

There are perhaps thousands of projects worldwide that use the Dublin Core for cataloging or to collect data from the Internet. The subjects range from cultural heritage and art to math and physics. Dublin Core is the basis for the Connexion System which OCLC has developed as its web-based cataloging system for all resources. Dublin Core is also the minimum shareable metadata set in the Open Archive Initiative-Protocol for Metadata Harvesting. While other sets can be used based on mutual agreement between the data provider and the harvester, every OAI-compliant provider must provide unqualified Dublin Core metadata.

3.2 Metadata Object Description Schema (MODS)

MODS is a schema for a bibliographic element set that is intended to support the interoperability of MARC records (especially MARC 21) with other bibliographic metadata schemes. It was developed by the Library of Congress for a variety of applications, particularly those related to library catalogs. It includes a subset of MARC fields, but it uses language-based tags rather than the traditional numeric ones used by MARC. MODS includes 19 top level elements which in some cases regroup the MARC elements. MODS is expressed in XML and is often used in conjunction with METS (see section 4.1.1 below) as a transfer format. MODS 3.0 was released in March 2004.

3.3 Global Information Locator Service (GILS)

GILS was developed by the U.S. government as a tool for enhancing public access to government information. Originally called the "Government Information Locator Service", GILS in various forms has been adopted by other governments and for international projects, leading to its current name, "Global Information Locator Service". International implementers of GILS include Australia, Germany, Singapore, and Hong Kong. GILS is also widely used with spatial and environmental clearinghouses implemented by countries and international organizations.

GILS specifies a profile of the Z39.50 protocol for distributed search and retrieval which is a common standard used in online library catalogs. It specifies the attributes (or the elements) that must be able to be searched in order for a system to be GILS compliant. However, organizations have specifically defined GILS elements for their own communities.

Since the purpose of GILS is to act as a locator service, GILS elements emphasize availability and distribution rather than description. Therefore, a GILS record may have elements such as the name and address of the distributor and information on ordering.

A U.S. Federal GILS Core Record For This Paper

Title: Metadata for Electronic Information Resources

Originator: Gail Hodge

Local Subject Term: Metadata

Abstract: Describes metadata standards for electronic libraries and related projects.

Purpose: To serve as an educational aid to librarians, information center managers and others involved in the dissemination and creation of electronic resources.

Availability:

Distributor :

Name: Information International Associates (IIa)

Street Address: 1009 Commerce Park Dr., Suite 150

City: Oak Ridge

State: TN

Country: USA

Zip Code: 37830

Telephone: 865-481-0388

Fax: 865-481-0390

Order Process: This paper is available without charge by writing to IIa at the address provided.

The original goal of GILS was to provide high-level locator records for government resources, both electronic and non-electronic. GILS records were intended to describe aggregates or collections such as catalogs, publishing services and databases. However, some organizations use GILS at the individual item (journal article or technical report) level.

3.4 Text Encoding Initiative (TEI) Header

The Text Encoding Initiative is an international project to develop guidelines for marking up electronic texts such as novels, plays, and poetry, primarily to support text analysis. As part of the mark-up a header portion has been defined, which includes metadata about the work. The TEI header, like the rest of the TEI, is defined as a Standard Generalized Mark-up Language Document Type Definition (SGML DTD).

The information in the TEI Header is similar to that captured in a library catalog. In fact, the TEI tag set can be mapped to and from MARC. In addition, elements are defined that record non-bibliographic information about the text itself, for example, how the text was transcribed or edited, what revisions have been made, and who performed the mark-up. All these metadata elements are important in text analysis and textual scholarship.

3.5 Encoded Archival Description (EAD)

Finding aids are important tools for resource description and discovery in archives and special collections of both physical and digital records. Finding aids differ from traditional library catalog records by being much longer, more narrative and explanatory, and hierarchical in their structure. The Encoded Archival Description (EAD) was developed as a way of marking up the data contained in a finding aid, so that it can be searched and displayed online. The EAD is particularly popular in academic libraries with large special collections and in archives. Users of EAD hope this scheme will encourage consistency and facilitate cross-archive searching. The EAD standard is maintained jointly by the Library of Congress and the Society of American Archivists.

Like the TEI Header, the EAD is defined as an SGML DTD. It begins with a header section that describes the finding aid itself (for example, who wrote it) which could be considered metadata about the metadata. It then describes the whole collection or record series and successively more detailed information about the contents of the collection. When the individual items being described exist in digital form, the EAD record can include pointers (digital identifiers) to the electronic information.

3.6 ONIX International

ONIX (Online Information Exchange) International is a metadata scheme developed by a number of book industry trade groups in the United States and Europe to support e-commerce. ONIX has elements for basic bibliographic, trade, evaluation and promotional information for books and e-books. This metadata standard is particularly valuable on Internet-based booksellers, such as Amazon.com. It supports the display of such online features as pictures of book covers, book review "snippets", and links to author biographies. Although initially focused on books, ONIX has been adapted to serial publications.

3.7 Content Standard for Digital Geospatial Metadata

Metadata schemes for datasets are particularly significant in disciplines where numeric and statistical data are primary resources. One of the most well developed element sets and content standards for data is the ISO Standard for Digital Geospatial Metadata (ISO 19115:2003). Geospatial datasets link data for a specific purpose to the latitude and longitude coordinates on the earth. These datasets are used in a wide variety of applications, including soil and land use studies, climatology and global change monitoring, remote sensing, and demographic and social science research.

The ISO Standard defines over 200 elements. The majority of these elements are optional in the standard, but they may be mandatory for specific implementations. Many national and local governments use the content standard, and it has become deeply embedded in Geospatial Information Systems (GIS). The standard forms the basis for the work of the Open GIS Consortium to provide for better interoperability among GIS applications.

3.8 Data Documentation Initiative (DDI)

The Data Documentation Initiative is a consortium of public and private sector organizations including major universities and the U.S. Bureau of the Census. The DDI's goal is to establish metadata standards for describing social science data sets. Included are elements such as the collection method, relevant software, and units of measure. A similar initiative within the U.S. Bureau of the Census involves metadata to describe questionnaires and other survey instruments.

3.9 Technical Metadata for Digital Still Images

The National Information Standards Organization in the United States has developed a data dictionary of technical elements for digital still images (July 2000, draft released for comment, February 2001). NISO realized that the focus of most cultural institutions had been on descriptive metadata, without any emphasis on the technical aspects of digital images that would be needed to adequately store and preserve them. The purpose of the standard is to facilitate the "development of applications to validate, manage, migrate and process images of enduring value." The emphasis is not only on current use of still images, but on the long-term provenance, preservation, and assessment for use and re-use.

The draft standard is quite extensive. The Basic Image Parameters section alone includes over 50 elements. For example, there are elements that describe the format, such as compression, MIMEtype and photometric interpretation. Elements related to the image's creation include the scanning agency and camera capture settings. The change history includes the processing agency and the processing software. There are additional elements such as spatial metrics, the colormap, the image width, and the image length.

4.0 METADATA INTEROPERABILITY

With so many metadata schemes, how will chaos be avoided? How can we ensure that systems that use different metadata schemes will be interoperable, in other words that information collected by one organization for a particular purpose can be exchanged, transferred or used by another organization for a different purpose. Practitioners cite metadata frameworks, crosswalks, and metadata registries as ways to achieve this interoperability. However, it should be noted that there has been little large scale testing of metadata interoperability.

4.1 Metadata Frameworks

A metadata framework is a reference model that provides a high-level, conceptual structure into which other metadata schemes can be placed. It also gives designers and developers a consistent, cross cutting terminology around which to discuss metadata for a particular purpose.

4.1.1 Metadata Encoding and Transmission Standard

The Metadata Encoding and Transmission Standard (METS) was developed by the Digital Library Federation and the Library of Congress for the management of digital library objects. METS uses a framework, described earlier in this paper, which defines metadata as descriptive, administrative or structural. The most significant contribution of METS is its emphasis on structural metadata. METS also adds a fourth component, a list of the files in the digital library object. The structural component of the METS scheme indicates how these files work together to form the digital library object. This structure information supports the management of the object by a digital library, and it facilitates the exchange of these objects among digital libraries.

METS provides an XML DTD that can point to metadata in other schemes by declaring the scheme that is being used. For example within the METS framework, Dublin Core elements could be used to describe a digital still image for resource discovery, and the technical elements from NISO's Draft Standard for Digital Still Images could be used to document the structural aspects of the image.

4.1.2 <indecs>

The Interoperability of Data in E-Commerce Systems (<indecs>) Framework is an international collaborative effort originally supported by the European Commission. It has developed a metadata framework, or a reference model, that supports the sharing of information about intellectual property rights in electronic commerce. In the basic model, people make "stuff", people use "stuff", and people make deals about "stuff."

Rather than develop a new metadata standard, <indecs> provides a framework for the various existing schemes to interact. For example, transactions related to music, journal articles or books could interchange information with one another. This framework has also been discussed as a way to allow the various groups (publishers, libraries and users) involved in access to electronic journal subscriptions to work within a consistent framework for interchange while maintaining the original metadata for their local applications.

In a significant move, the <indecs> framework has been adopted by the MP3 standards group working on standards for multimedia including intellectual property. In the MP3 context, the <indecs> framework is known as Contecs:DD.

4.1.3 Open Archive Initiative

The Open Archive Initiative (OAI) began as a project to provide consistent access across the numerous e-print services created by government and academia in the mid-1990s. However, the OAI has proven to be generally applicable for other types of electronic resources. The objective of the OAI is to create a low barrier to implementation, so OAI has only a few metadata elements based on the Dublin Core. Communities can extend the minimal set as needed.

To be OAI-compliant, the archive exposes the OAI metadata set by crosswalking its native metadata format to that of the OAI. This file is exposed and then harvested into a central repository. The software for implementing an OAI compliant archive is freely available. Recent developments include tools for searching OAI repositories, some of which focus on the needs of specific communities.

4.2 Metadata Crosswalks

Metadata crosswalks map the elements, semantics and syntax from one metadata scheme to those of another. A crosswalk allows metadata created by one community to be used by another community with a different metadata standard. The degree to which crosswalks are successful depends on the similarity of the two schemes. The mapping of schemes with fewer elements (less granular or atomic) to those with more elements (more granular or atomic) is problematic. Despite similarity at the semantic level, the crosswalk can be difficult if the content rules differ from the original scheme to the target scheme even if the definitions of the elements are similar.

While these crosswalks are key to interoperability, they are also labor intensive to develop and maintain. However, crosswalks are important for virtual libraries and subject gateways that collect or search resources from a variety of sources and treat them as a whole collection.

4.3 Metadata Registries

Registries are another tool for exchanging metadata. They provide information about the definition, origin, source, and location of the scheme, usage profile, element set, and/or authority files for element values. A registry maps one scheme to another so that both humans and computers can understand how they might integrate, and registries can also document rules for transforming content for an element in one system to the content required for an equivalent element in another. The DESIRE (Development of a European Service for Information on Research and Education) Project funded by the European Commission has developed a prototype of such a registry based on the ISO standard for defining data elements (ISO/IEC 11179). The Dublin Core Metadata Initiative has also developed a registry for the Dublin Core elements.

Registries are particularly useful in specific disciplines or industries such as health care, aeronautics, or environmental science, where they can be used to make the contents of resources more easily integrated. A good example is the U.S. Environmental Protection Agency's Environmental Data Registry which provides information about thousands of data elements used in current and legacy EPA databases. The EDR metadata registry provides an integrating resource for legacy data, acts as a look-up tool for designers of new databases, and documents each data element. The European Environment Agency has developed a similar registry which is available as open source software on the Web.

5.0 METADATA CREATION

Metadata is extremely important for the discovery and management of digital resources. However, there are major issues related to the cost and time involved in creating this metadata. A variety of methods are used for creating metadata ranging from manual creation to metadata creators/editors, metadata generators and metadata harvesters.

5.1 Manually Created Metadata

Who creates metadata? The answer to this question varies by discipline, the electronic information being described, the tools available, and the expected outcome. In the case of descriptive metadata, originators may provide some level of metadata creation. This is particularly true in the documentation of datasets where the originator has significant understanding of the rationale for the dataset, the way the data was collected, and the uses to which it could be put, and where there is little if any textual information that a cataloger could use. In other cases, projects have found that it is necessary to have metadata catalogers or librarians create the metadata or at least review the metadata created by the originators, because the originators do not have the time or the skills to create adequate metadata.

The cost of creating metadata for the burgeoning number of electronic resources has led to the development of two types of tools – metadata generators and metadata creators/editors. Creators/editors support the manual creation and editing of metadata. Metadata generators automatically create a metadata record based on the original source.

5.2 Metadata Creators/Editors

Metadata creators/editors, both commercial and proprietary, address the need for speed and quality. Many tools support validation rules and pick lists based on authority files or controlled domains, including controlled vocabularies and thesauri. Templates may be provided and customized to stream line the data entry process.

There are several metadata creators/editors for the Dublin Core. The Nordic Web provides metadata creation software and Dublin Core to MARC conversion software, which is free within the European Union. MetaWeb from Australia has developed a metadata editor called “Reggie.”

There are a number of FGDC-compliant metadata creation tools, including Metamaker, which was developed by the U.S. Geological Survey, Biological Resources Discipline. Some FGDC-compliant products have been developed by geographic information system (GIS) vendors to support the documentation of information created or stored within their products. While many of these systems are proprietary, the Open GIS Consortium has developed standards for open metadata tools.

5.3 Metadata Generators

The program DC.doc from the UK Online Library Network (UKOLN) analyzes a Web site (indicated by a URL that the user provides) and creates a Dublin Core record. The proposed content is displayed back to the user in a Dublin Core template. The user can modify the content. The DC fields can be returned to the Web site as metatags or stored in a separate file. Similar programs are available from The Nordic Web Project, OCLC’s Connexion system for shared cataloging of Web resources, and Australia’s MetaWeb.

Of course, the content of metadata generators is only as good as the content of the originating Web site. None of these tools provide 100% automatic metadata generation, particularly if high quality content is desired. Users often use the software simply for a handy Dublin Core template. However, efforts are underway at Syracuse University with funding from the National Science Foundation's National Science Digital Library to improve the results of automatic metadata generation.

5.4 Metadata Harvesters

Metadata harvesters are programs or scripts that capture metadata from other metadata sources. They are slightly different from metadata generators in that they access sets of metadata that are already created. The most well known example of a metadata harvester is the Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH). This script which is available as open source software is designed to identify repositories of OAI-compliant metadata. Using the OAI-PMH organizations can create aggregated repositories of metadata for distributed digital objects. This metadata can be merged with the organization's own metadata. Aggregations can be built around subjects, library consortia or regional collaborations. The minimum requirement is the exposure of the unqualified Dublin Core Metadata element set.

A similar approach is used by OCLC's Connexion system. Based on the principal of shared cataloging, this system allows libraries to share the Dublin Core-based metadata records that have been created by other libraries. As with other shared cataloging activities, the basic metadata structure is well controlled but content standards must be agreed upon and consistently implemented or levels of quality will vary.

6.0 CONTROLLED VOCABULARIES AND METADATA

The use of controlled vocabularies is becoming increasingly important as a tool for metadata creation and access. This is particularly true as more information managers realize the problems that arise from free text searching or the use of uncontrolled keywords.

Most metadata schemes do not dictate the use of a particular controlled vocabulary when entering the contents of elements that describe what the resource is about. However, use of controlled vocabularies is encouraged, particularly within a subject domain. Many metadata schemes allow controlled vocabularies to be defined within the syntax.

A variety of controlled vocabulary systems are being used for indexing electronic resources. These include traditional library schemes such as the Library of Congress Subject Headings and the Dewey Decimal Classification, specific domain-oriented thesauri or classification schemes, and locally created lists of frequently used or important terms. The tool suite required to use existing controlled vocabulary schemes in the Internet environment is a major research area for OCLC.

Individual projects may specify the controlled vocabularies to be used. For example, the National Biological Information Infrastructure, which uses the Biological Profile for the FGDC Geospatial Content Standard, specifies the controlled vocabulary to be used. Cambridge Scientific Abstracts, as a partner of the National Biological Information Infrastructure (NBII), has developed a Biocomplexity Thesaurus. The terms in the thesaurus will be used in the NBII metadata to tag electronic resources across the NBII subject and geographic nodes. The thesaurus will also be used to select terms for the more traditional bibliographic indexing in CSA's Biocomplexity Database, which is searchable through the NBII Web site. The NBII portal will use the terms to create collections of information based on a user's personal preferences. The NBII's Biological Profile of

the FGDC Metadata Content Standard also specifies the use of the Integrated Taxonomic Information System as the authority file for completing the biological taxonomic classification elements within the metadata record.

Controlled term lists have been developed by many of the U.S. states using the Global Information Locator Service. These include terms that describe the major services and products provided by states to their citizens, to state employees, or to other governments, whether state, local or national.

Unfortunately, the use of individual controlled vocabulary schemes does not significantly improve searching across the breadth of Internet resources or when the user is searching outside his or her area of expertise. A group called the Networked Knowledge Organization Systems/Services (NKOS), an ad hoc group from public and private sector organizations in ten countries, has been discussing the issues related to providing generally applicable knowledge organization services (KOS) via the Internet. The group defines KOSs to include authority files, thesauri, gazetteers, ontologies, topic maps, taxonomies, subject headings, and other types of schemes intended to organize digital objects. NKOS has been discussing protocols for the use of KOSs via the Internet, and has developed a set of metadata elements to describe KOSs and their characteristics and behavior. This metadata could be used as part of a registry of KOSs or as metatag information embedded in header information for a Web-based KOS.

Another approach to making controlled vocabularies more generally available as Internet tools is the development of terminology Web services. A Web service uses specific standardized protocols to create modules that can be used and re-used in a variety of applications over the Web. For example, the U.S. National Agricultural Library has developed its Web-enabled Agricultural Thesaurus as a Web service. Functionality includes locating a term, navigating the thesaurus, and selecting the term and various types of related terms. This functionality can then be incorporated by any other system wanting to use the Agricultural Thesaurus as the basis for searching or browsing its content.

In a similar initiative, a Z39.50 profile for thesauri has been developed. The profile provides a high level, abstract representation for navigating a thesaurus. In addition to providing thesaurus search capabilities within the realm of Z39.50 (which includes the GILS initiatives and many of the initiatives that use the FGDC content standard) an appendix to the profile provides an XML DTD for thesauri that could be used by other protocols.

Web services and other networked applications of controlled vocabularies help to support the development of a Semantic Web, a major activity of the World Wide Web Consortium. The goal of this initiative is to provide the Web with an “understanding” of concepts in order to result in better machine processing of text and provision of services. The basis for the Semantic Web is a knowledge representation that is much richer than that reflected in standard thesauri or classification schemes. These ontologies, semantic networks or topic maps encode more specific relationships between concepts. For example, instead of just labeling leg as a narrower term to body part, the relationship would be specifically identified as “whole-part”. In a knowledge organization system concerned with the environment and human health, the relationship between mosquito and West Nile Virus might be “carrier (or vector)- disease”. The same term, mosquito would have another relationship to the term Insect as its higher level biological taxonomy relationship. Having more explicit relationships allows for better disambiguation of results and the building of rules and assumptions into information retrieval systems.

The SWAD-E group in Europe has developed SKOS Core 1.0, an RDF schema for thesauri and other similar knowledge organization systems. It is intended to serve as the basis for moving traditional knowledge

organization systems into formats that are more appropriate for the Semantic Web even though the rich semantic relationships may need to be enhanced manually or through additional programming.

7.0 CONCLUSIONS

Metadata schemes have been developed for a variety of purposes – resource discovery, location, collection organization and management, administration, rights management, technical reproducibility and preservation. However, because the needs of resource types and user communities differ, many schemes have been developed, along with specific extensions and profiles. Metadata standards and interoperability remain key issues. In order to increase the use of metadata, systems need to be developed that support metadata creation at the same time that the resource is created. Larger testbeds of metadata and search engines that take more advantage of metadata that has been created must be developed. Communities of practice need to develop content standards and to look for common areas of interest in order to support access to information across communities. Most importantly, creators of electronic resources must be made aware of the importance of metadata for the short as well as the long-term use of their contributions to the world of electronic information resources.

8.0 SELECTED RESOURCES ON METADATA, FRAMEWORKS AND RELATED STANDARDS[1]

General Resources about Metadata

Distributed Systems Technology Centre. (2000). Metadata Schema Registry (Australia). Retrieved April 21, 2004 from the Metadata Schema Registry Web site: metadata.net/

Hodge, G. (2001). Metadata Made Simpler: A Guide for Libraries Retrieved April 21, 2004 from the National Information Standards Organization Web site: www.niso.org/news/Metadata_simpler.pdf

International Federation of Library Associations and Institutions (IFLA). (2002). Digital Libraries: Metadata Resources. Retrieved April 21, 2004 from the International Federation of Library Associations and Institutions Web site: www.ifla.org/II/metadata.htm

Metadata Information Clearinghouse (Interactive). (1999). Retrieved April 21, 2004 from the Metadata Information Clearinghouse Web site: www.metadatainformation.org/

Schwartz, C. (2002). Metadata Portals & Multi-standard Projects. Retrieved April 21, 2004 from the Simmons College Web site: web.simmons.edu/~schwartz/meta.html

UK Online Library Network (UKOLN). (2002). Metadata Resources. Retrieved April 21, 2004 from the UK Online Library Network Web site: www.ukoln.ac.uk/metadata/resources

Selected Metadata Schemes and Frameworks

Data Documentation Initiative: DDI. Retrieved April 21, 2004 from the University of Michigan Web site: www.icpsr.umich.edu/DDI/

[1] Inclusion in this list does not constitute endorsement by Information International Associates or the U.S. Geological Survey.

DESIRE (Development of a European Services for Information on Research and Education). (2002). Metadata Registry. Retrieved April 21, 2004 from the UK Online Library Network Web site: desire.ukoln.ac.uk/registry/

Dublin Core Metadata Initiative. (2002). Retrieved April 21, 2004 from the OCLC Web site: purl.oclc.org/metadata/dublin_core/

Ellerman, Castedo. (1997). Channel Definition Format (CDF). Retrieved April 21, 2004 from the WWW Consortium Web site: www.w3.org/TR/NOTE-CDFsubmit.html

Encoded Archival Description (EAD). (2001). Retrieved April 21, 2004 from the Library of Congress Web site: lcweb.loc.gov/ead/

FGDC Content Standard for Digital Geospatial Metadata (CSDGM). (2001). Retrieved April 21, 2004 from the Federal Geographic Data Committee Web site: www.fgdc.gov/metadata/contstan.html

GEM (Gateway to Educational Materials) Element Set & Profile(s) Workbench. (2002). Retrieved April 21, 2004 from the GEM Web site: www.geminfo.org/Workbench/Metadata/index.html

Global Information Locator Service (GILS). Retrieved April 21, 2004 from the Global Information Locator Service Web site: www.gils.net

IMS Global Learning Consortium, Inc. (2002). Learning Resource Meta-data Specification. Retrieved April 21, 2004 from the IMS Global Learning Consortium Web site: http://www.imsproject.org/metadata/

<indecs> interoperability of data in e-commerce systems. Retrieved April 21, 2004 from the <indecs> Web site: www.indecs.org/

METS: Metadata Encoding and Transmission Standard. (2001). Retrieved April 21, 2004 from the Library of Congress Web site: www.loc.gov/standards/mets/

MODS: Metadata Object Description Schema. (2004). Retrieved June 18, 2004 from the Library of Congress Web site: www.loc.gov/standards/mods/

OCLC/RLG Working Group on Preservation Metadata. (2001). Preservation Metadata and the OAIS Information Model: A Framework to Support the Preservation of Digital Objects. Retrieved June 4, 2004 PREMIS (Preservation Metadata: Implementation Strategies). Retrieved June 4, 2004 from the OCLC Web site: http://www.oclc.org/research/projects/pmwg/

ONIX (Online Information Exchange). Retrieved April 21, 2004 from the Editeur Web site: www.editeur.org/

PREMIS (Preservation Metadata: Implementation Strategies). Retrieved June 4, 2004 from the OCLC Web site: http://www.oclc.org/research/projects/pmwg/

Technical Metadata for Digital Still Images. (2001). Retrieved April 21, 2004 from the National Information Standards Organization Web site: www.niso.org/committees/committee.au.html

TEI Consortium. Text Encoding Initiative. (2002). Retrieved April 21, 2004 from the TEI Web site: www.tei-c.org/

Visual Resources Association Data Standards Committee. VRA Core Categories, Version 3.0. (2002). Retrieved April 21, 2004 from the VRA Web site: www.vraweb.org/vracore3.htm

Metadata Crosswalks

Day, M. (2001). Metadata: Mapping between Metadata Formats (comprehensive list of mappings to and from all major formats including national versions of MARC). Retrieved April 21, 2004 from the UK Online Library Network Web site: www.ukoln.ac.uk/metadata/interoperability/

St. Pierre, M. and W. La Plant. (1998) Issues in Crosswalking Content Metadata Standards. Retrieved June 4, 2002 from the National Information Standards Organization Web site: http://www.niso.org/press/whitepapers/crsswalk.html

Metadata Tools

BlueAngel Technologies (MetaStar). Retrieved April 21, 2004 from the BlueAngel Technologies Web site: www.blueangeltech.com/

Distributed Systems Technology Centre. (1999). MetaWeb Project (Australia). Retrieved April 21, 2004 from the Distributed Systems Technology Centre Web site: www.dstc.edu.au/RDU/MetaWeb

Dublin Core Metadata Initiative. (2002). Tools and Software. Retrieved April 21, 2004 from the Dublin Core Web site: dublincore.org/tools/

Federal Geographic Data Committee. Metadata Tools. Retrieved April 21, 2004 from the Federal Geographic Data Committee Web site: www.fgdc.gov/metadata/metatool.html

Intergraph Spatial Metadata Management System. (2002). Retrieved April 21, 2004 from the Intergraph Web site: www.intergraph.com/gis/smms//

Meta Matters. Retrieved April 21, 2004 from the National Library of Australia Web site: http://dcanzorg.ozstaging.com/mb.aspx

Metadata: UKOLN Software Tools (comprehensive list of tools for a variety of standards including Dublin Core, GILS and IMS). Retrieved April 21, 2004 from the UK Online Library Network Web site: www.ukoln.ac.uk/metadata/software-tools/

MetaPackager. (2004). Retrieved June 3, 2004 from the HiSoftware Web site: http://www.hisoftware.com/xmlp/metapackager.htm

Nordic Metadata Projects. (2000). Retrieved April 21, 2004 from the University of Helsinki Library Web site: www.lib.helsinki.fi/meta/

Related Initiatives

Australian Government Locator Service Metadata Standard. (2000). Retrieved June 4, 2004 from the National Archives of Australia Web site: http://www.agls.gov.au/

CORC (Cooperative Online Resource Catalog). (2002). Retrieved April 21, 2004 from the OCLC Web site: www.oclc.org/corc/about/

CrossRef. (2000). Retrieved April 21, 2004 from the CrossRef Web site: www.crossref.org

National Biological Information Infrastructure (U.S.). (2001). Retrieved April 21, 2004 from the NBII Web site: www.nbii.gov

National Spatial Data Infrastructure (U.S.) (2002). Retrieved April 21, 2004 from the Federal Geographic Data Committee Web site: www.fgdc.gov/nsdi/nsdi.html

Networked Knowledge Organization Systems/Services (NKOS). (2002). Retrieved April 21, 2004 from the School of Library and Information Science, Kent State University Web site: nkos.slis.kent.edu/

Open Archives Initiative. (2002). Retrieved April 21, 2004 from the OAI Web site: www.openarchives.org

W3C Technology and Society Domain: Semantic Web Activities. Retrieved June 4, 2004 from the WWW Consortium Web site: http://www.w3.org/2001/sw/

Related Standards and Best Practices

Corporation for National Research Initiatives. Handle System®. (2002). Retrieved April 21, 2004 from the Handle Web site: www.handle.net/

Extensible Markup Language (XML). (2002). Retrieved April 21, 2004 from the WWW Consortium Web site: www.w3.org/XML/

International DOI Foundation. DOI: Digital Object Identifier System. (2002). Retrieved April 21, 2004 from the International DOI Foundation Web Site: www.doi.org/

PURL (Persistent URL). (2002). Retrieved April 21, 2004 from the PURL Web site: purl.org

Resource Description Framework. (2002). Retrieved April 21, 2004 from the WWW Consortium Web site: www.w3.org/RDF/

Taylor, M. Zthes: a Z39.50 Profile for Thesaurus Navigation, Version 0.4. (2000). Retrieved April 21, 2004 from the Library of Congress Web site: lcweb.loc.gov/z3950/agency/profiles/zthes-04.html

W3C: World Wide Web Consortium. (2002). Retrieved April 21, 2004 from the WWW Consortium Web site: www.w3.org/

Z39.50. (2002). Retrieved April 21, 2004 from the Library of Congress Web site: http://www.loc.gov/z3950/agency/

Preservation of and Permanent Access to Electronic Information Resources

Gail Hodge
Information International Associates, Inc.
312 Walnut Place
Havertown, Pennsylvania 19083
USA

gailhodge@aol.com

ABSTRACT

The rapid growth in the creation and dissemination of electronic information has emphasized the digital environment's speed and ease of dissemination with little regard for its long-term preservation and access. To some extent, electronic libraries, that is those libraries that are moving toward provision of materials in electronic form, have been swept up in this attitude as well. Electronic information includes a variety of object types such as electronic journals, e-books, databases, data sets, reference works, and web sites, which are born digital or which have their primary version in digital form.

But, electronic information is fragile in ways that traditional paper-based information is not. Electronic information is more easily corrupted or altered, intentionally or unintentionally, without the ability to recognize that the corruption has occurred. Digital storage media have unknown life spans. Some formats, such as multimedia, are so closely linked to the software and hardware technologies that they cannot be used outside these proprietary environments. Aggravating this situation is the fact that the time between creation and preservation is shrinking, because technological advances are occurring so quickly.

The Open Archival Information System (OAIS) Reference Model provides a framework for discussing the key areas that impact on digital preservation – the creation of the electronic information, the acquisition of and policies surrounding the archiving of resources, preservation formats, preservation planning including issues of migration versus emulation, and long-term access to the archive's contents.

Many projects, worldwide, have contributed to the growing collection of best practices and standards. The numerous stakeholder groups involved in preservation of electronic resources, including creators (authors), publishers, librarians and archivists, and third-party service providers, are working more closely to build a cohesive and sustainable response to the issues. An issue of continuing stakeholder interest is the economic model(s) that will provide ongoing support to electronic preservation.

Despite the remaining issues, local institutions managing electronic libraries can become involved. They are encouraged to monitor developments and projects in the field, to raise awareness of the need for preservation within their institutions, to consider preservation and long-term access issues when negotiating licenses for electronic resources, and to look for opportunities to begin small projects at the local level.

1.0 BACKGROUND

Major activities have been underway in digital archiving and preservation since the early 1990s. In an effort to quickly focus on the current state of the practice and research, this section provides a definition of key terms that will be used throughout the paper and introduces important projects that are used as examples.

1.1 Definition of Terms

Key terms used throughout this paper are defined below. In some cases, these definitions are for consistency within the presentation and are not indicative of general consensus within the community.

Born digital – materials that are created in bits and bytes rather than being digitized from paper or other analog medium

Digital archiving – storing the digital information (e.g., creating an institutional repository or digital archive)

Digital preservation – keeping the bits and bytes safe and unaltered for a long period of time

Digitization – converting materials in non-digital form (analog) such as paper, to digital form

Emulation – running old products by recreating the environment of the old hardware and software without actually using the old hardware and software

Migration – moving a digital product from one version of a program, operating system or hardware environment to another over time

Permanent or Long-term access – the ability to use a preserved object long after its initial preservation

Recapturing – copying the content from the original resource again in order to ensure that changes made to the resources are incorporated in the archival version

Refreshing – moving a digital object to a new instance of the same storage medium, retaining the same operating system and hardware environment

1.2 Outline of Major Projects and Operational Systems

There are several major projects in digital preservation which can serve as examples. (Good sources for information about such projects include the PADI (Preserving Access to Digital Information) Web site from the National Library of Australia (NLA 2003), the joint Web site from PADI and the Digital Preservation Coalition (DPC/PADI 2004), and RLG's *DigiNews* newsletter.) This section briefly describes these major research projects and operational systems, since they are used in the following sections of the paper.

CAMiLEON, (Creative Archiving at Michigan and Leeds: Emulating the Old on the New) a joint project of the University of Michigan and the University of Leeds, conducted analysis and testing to determine if emulation is a viable technical strategy for preservation (CAMiLEON 2001).

Cedars (CURL Exemplars in Digital Archiving) was sponsored by the Joint Information Systems Committee in the UK. It was established to determine the feasibility of distributed digital archives. The first implementation included the three institutions in the Consortium of University Research Libraries. In order to prove scalability, Cedars incorporated several other test sites. This project was completed in 2002 with a series of guideline documents (Cedars n.d.).

DIAS (Digital Information Archive System) was developed by IBM for the Dutch National Library. It is based on the OAIS Reference Model and outcomes of the NEDLIB project of the European Union which was completed in early 2001 (NEDLIB 2001). It is considered to be the first commercially available operational archiving system for electronic journals. It is particularly geared toward the needs of national libraries with

legal deposit responsibilities. The current implementation deals primarily with the deposit and tracking aspects. Major issues such as preservation and long-term access are still being investigated.

Digital Preservation Coalition (DPC) is the umbrella organization for the UK preservation efforts. It has incorporated many of the organizations and lessons learned from projects such as Cedars. Most recently it has been instrumental in promoting the incorporation of non-print materials in legal deposit in the UK (Digital Preservation Coalition 2002).

DSpace is a suite of tools for developing an institutional repository to archive various digital objects. There are several implementations, most notably at the Massachusetts Institute of Technology Libraries. The current focus is on archiving, but there is a planned component for preservation and for the creation of federated repositories.

ERPANET (Electronic Resources Preservation and Access Network) is funded by the European Commission to provide a knowledge base and advice to all sectors on issues of archiving and preservation of electronic resources. ERPANET is best known for its workshops, including those related to archiving various types of digital objects and those related to electronic records (ERPANET 2004).

EVA is a project of the National Library of Finland at the University of Helsinki, which uses a series of automatic tools including robots, harvesters, and metadata creation tools to support its goal of capturing electronic network publications of Finland (Lounamaa and Salonharju 1999).

InterPARES (International Research on Permanent Authentic Records in Electronic Systems) is a global project among archiving institutions, including regional consortia for Asia and Europe. The project's goals are to develop best practices related to the creation, preservation and long-term access to *authentic* electronic records. In its second phase, InterPARES2 is investigating issues such as multimedia and dynamic content preservation (InterPARES n.d.).

Kulturaw3 is a project of the Royal Library of Sweden to capture the cultural heritage that is being published via the Internet (Royal Library n.d.).

JSTOR, originally funded by the Andrew J. Mellon Foundation, is now a non-profit organization that archives back issues of journals for publishers by digitizing them. It is just beginning to deal with current journal issues that are in electronic form (JSTOR 2004).

LOCKSS (Lots of Copies Keep Stuff Safe) is a project of the Stanford University Library, its publishing arm, HighWire Press, and several other libraries to develop a system for redundant archives. Its major contribution is an infrastructure for keeping redundant archives synchronized. There is a special project called LOCKSS-DOCS, which is focused on the preservation of U.S. government documents (LOCKSS n.d.).

NDIIPP (National Digital Information Infrastructure and Preservation Program) is a program at the Library of Congress to develop an infrastructure for digital preservation within the U.S. It is viewed as a joint partnership with content owners ranging from traditional publishers to multimedia providers. In cooperation with the National Science Foundation, NDIIPP has developed a research agenda. In its current phase, it is developing partnerships across the stakeholders and promoting solutions to the research agenda through a series of grant awards.

OCLC Digital Archive is a service of OCLC that grew out of its electronic journals' project. In this service OCLC acts as a trusted third party archive receiving deposits of electronic journals into its repository.

It provides several levels of access (continuous or just in case) and controls access rights so that a library can access only the issues equating to the period for which it had a license (OCLC 2004a).

PANDORA (Preserving and Accessing Networked DOcumentary Resources of Australia), a project of the National Library of Australia, captures the Web-based cultural heritage of Australia. It involves capturing content, creating metadata, and making arrangements with rights holders. A federated approach includes the libraries in all the Australian states. (PANDORA n.d.) PANDAS is the operational system that supports PANDORA.

PREservation Metadata: Implementation Strategies (PREMIS) is a joint project of OCLC, the Research Libraries Group (RLG), and others, to establish a standard metadata element set for preservation and to identify best practices related to implementation (OCLC 2004b).

2.0 A FRAMEWORK FOR ARCHIVING AND PRESERVATION

It is valuable to discuss archiving and preservation within a framework. The framework used in this paper is provided by a reference model, which is used extensively throughout the digital preservation community. The Open Archival Information System Reference Model (CCSDS 2001) provides high level data and functional models and a consistent terminology for discussing preservation. The reference model was originally developed by the Consultative Committee on Space Data Systems (CCSDS) to support the archiving of data among the major space agencies. However, it has become the de facto standard for the development of digital archives. It is used by most major projects including those in Australia, the United Kingdom, the Netherlands, and the United States. The OAIS Reference Model became a formal ISO standard in June 2002.

In its simplest form the OAIS looks like this (Fig. 1):

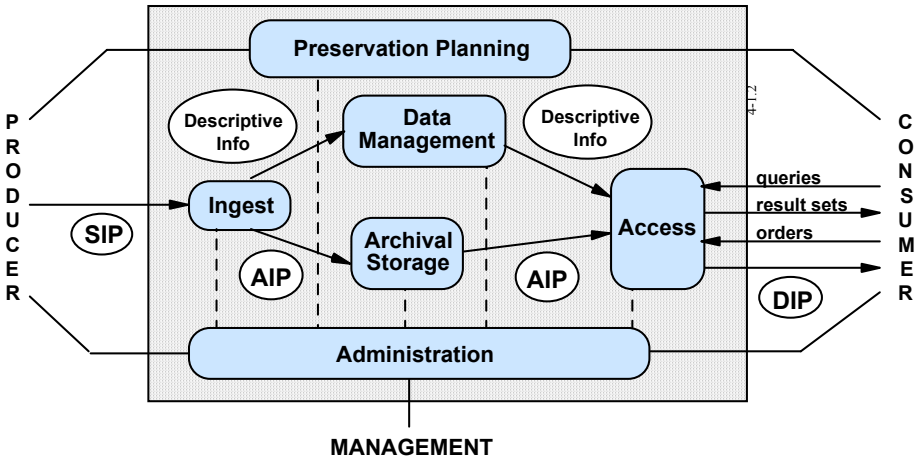


Figure 1: Open Archival Information System.

Source: Consultative Committee on Space Data Systems (used with permission)

SIP – Submission Information Packet (what is submitted or acquired from the producer)
AIP – Archival Information Packet (the object that is archived)
DIP – Dissemination Information Packet (the object that is distributed based on access requests)
Descriptive Info – Metadata

2.1 Production and Creation of Electronic Information

Preservation and permanent access begin outside the purview of the archive with the producer or the creator of the electronic resource. This is where long-term archiving and preservation must begin. Information that is born digital may be lost if the producer is unaware of the importance of preservation. Practices used when electronic information is produced will impact the ease with which the information can be digitally archived and preserved.

Several key practices are emerging involving the producers of electronic information. First, the archiving and preservation process is made more efficient when attention is paid to issues of consistency, format, standardization and metadata description before the material is considered for archiving. By limiting the format and layout of certain types of resources, archiving is made easier. This is, of course, easier for a small institution or a single company to enforce than for a national archive or library. In the latter cases, they are faced with a wide variety of formats that must be ingested, managed and preserved.

In the case of more formally published materials, such as electronic journals, efforts are underway to determine standards that will facilitate archiving, long-term preservation and permanent access. The Andrew J. Mellon Foundation has funded a study of the electronic journal mark-up practices of several publishers. The study concluded that a single SGML document type definition (DTD) or XML schema can be developed to support the archiving of electronic journals from different subject disciplines and from different publishers with some loss of special features (Inera, Inc. 2001). Such standardization is considered key to efficient archiving and preservation of electronic journals by third-party vendors. The DTD developed by PubMed Central for deposit of biomedical journals is being considered as a generalizable model for all journals. The Archiving and Interchange DTD Suite is based on an analysis of all the major DTDs that were being used for journal literature, regardless of the discipline. The suite is a set of XML building blocks or modules from which any number of DTDs can be created for a variety of purposes including archiving. Using the Suite, NLM created a Journal Archiving and Interchange DTD as the foundation for the PubMed Central archive. In addition, a more restrictive Journal Publishing DTD has been released which can be used by a journal to mark up its content in XML for submission to PubMed Central. Several publishers and projects, such as JSTOR, the Public Library of Science, High Wire Press and CSIRO, are analyzing or planning to use the Journal Publishing DTD (Beck, 2003).

In the case of less formally published material such as web sites, the creator may be involved in assessing the long-term value of the information. In lieu of other assessment factors, the creator's estimate of the long-term value of the information may be a good indication of the value that will be placed on it by members of its designated community or audience in the future. The Preservation Office at the National Library of Medicine has implemented a "permanence rating system" (Byrnes 2000). The rating is based on three factors: integrity, persistent location, and constancy of content. These factors have been combined into a scheme that can be applied to any electronic resource. At the present time, the ratings are being applied to NLM's internal Web sites, and guidelines have been developed to assist creators in assigning the ratings to their sites. This information will be used to manage the ongoing preservation activities and to alert users about a Web site's long-term stability.

Another aspect of the creator's involvement in preservation is the creation of metadata. The best practice is for metadata to be created prior to incorporation into the archive, i.e., at the producer stage. However, most of the metadata continues to be created "by hand" and after-the-fact. Unfortunately, metadata creation is not sufficiently incorporated into the tools for the creation of most objects to rely on the creation process alone. However, as standards groups and vendors move to incorporate XML and other architectures into software products, such as word processors, the creation of metadata should become easier and more automatic.

2.2 Ingest: Acquisition and Collection Development

The first function to be performed by the archive is acquisition and collection development. This is the stage in which the created object is "incorporated" physically or virtually into the archive. In the terminology of the reference model, this is called "Ingest". There are two main aspects to the acquisition of electronic information for archiving – collection policies and gathering procedures.

2.2.1 Collection Policies

Just as in the paper environment, there is more material that could be archived than there are resources with which to accomplish it. Guidelines are needed to tailor the collection policies to the needs of a particular organization and to establish the boundaries in a situation where the responsibility for archiving among the stakeholders is still unregulated. The collection policies answer questions such as what should be archived, what is the extent of a digital object, should the links that point from the object to be archived to other objects also be archived, and how often should the content of an archived site be recaptured?

2.2.1.1 Selecting What to Archive

In the network environment where any individual can be a publisher, the publishing process does not always provide the screening and selection at the manuscript stage on which traditional archiving practices have relied. Therefore, libraries are left with a larger burden of selection responsibility to ensure that publications of lasting cultural and research value are preserved (National Library of Canada 1998).

The scope of NLA's PANDORA (Preserving and Accessing Networked DOcumentary Resources of Australia) Project is to preserve Australian Internet publishing. The NLA has formulated guidelines for the *Selection of Online Australian Publications Intended for Preservation by the National Library of Australia* (NLA n.d.). Scholarly publications of national significance and those of current and long term research value are archived comprehensively. Other items are archived on a selective basis "to provide a broad cultural snapshot of how Australians are using the Internet to disseminate information, express opinions, lobby, and publish their creative work." The National Library of Canada has written similar guidelines (National Library of Canada 1998). The broadest guidelines for Collection Management are provided in a document from the Cedars Project (Weinberger 2000). The most comprehensive analysis of such guidelines is in the *Digital Preservation Handbook*, which is based on the combined lessons learned of all the major projects (Beagrie and Jones 2001).

Even the Internet Archive (Internet Archive n.d.), which considers the capture of the entire contents of the Internet as its mandate, has established limitations. The sites selected do not include those that are "off-limits," because they are behind firewalls, require passwords to access, are hidden within Web-accessible databases, or require payment.

The major lesson from efforts to develop selection guidelines is the importance of creating such a document in order to set the scope, develop a common understanding, and inform the users now and in the future what they can expect from the archive.

2.2.1.2 Determining Extent

Once the site has been selected for inclusion, it is necessary to address the issue of extent. What is the extent or the boundary of a particular digital work, especially when capturing a complex Web site? Is it a "home page" and all the pages underneath it, or are the units to be archived (and cataloged) at a more specific level?

The PANDORA (NLA/PANDORA) project in Australia evaluates both the higher and lower site pages to determine which pages form a cohesive unit for purposes of preservation, cataloging, and long-term access. While preference is given to breaking down large sites into components, the final decisions about extent depend upon which pages cluster together to form a stand-alone unit that conveys valuable information. Each individual component must meet PANDORA's initial selection guidelines.

2.2.1.3 Archiving Links

The extensive use of links in electronic publications raises the question of whether these links and their contents should be archived along with the original site. The answer to this question by any particular project will depend on the purpose of the archiving, the anticipated stability of the links, and the degree to which they contribute to the overall information value of the site.

Most organizations archive the URLs (Uniform Resource Locators) or other identifiers for the links and not the content of the linked pages, citing problems with the instability of links. Some projects have established variants on this approach. For example, PANDORA's decision to archive the content of linked objects is based on its selection guidelines; the content of the linked site is captured only if it meets the same selection criteria as other sites. The National Library of Canada captures the text of a linked object as long as it is on the same server as the object that is being archived, because these intra-server links have proven to be more stable than external links. The American Institute of Physics (AIP) points to the content of a linked reference if it is an item in AIP's archive of publications or supplemental material.

Elsevier cites a technology-related problem as the main reason it does not archive links (Hunter 2002). Elsevier's links are created on the fly, so there is no URL or live page to capture. Similar problems exist when trying to capture pages that are active server pages or those that are created out of a database, portal system, or content management system.

The American Astronomical Society (AAS) has perhaps the most comprehensive approach to the archiving of links. The AAS maintains all links to documents and supporting materials based on collaboration among the various astronomical societies, researchers, universities and government agencies involved in this specific domain. Each organization archives its own publications, retaining all links and access to the full text of all other links. In the future, similar levels of cooperation may be achieved in other subject domains or by publisher collaborations such as CrossRef.

2.2.1.4 Recapturing the Archived Contents

In cases where the site selected for archiving is updated periodically, recapturing the object is necessary. This would be the case for an electronic journal that publishes each article online as it becomes available or for a preprint service that allows the author to modify the content of the preprint as it proceeds through the review process.

When making decisions about recapturing the content of an archived site, a balance must be struck between the completeness and currency of the archive and the burden on the system resources. PANDORA allocates a gathering schedule to each "publication" in its automatic harvesting program. The options include on/off, weekly, monthly, quarterly, half-yearly, every nine months, or annually. The selection is dependent on the degree of change expected and the overall stability of the site. When making decisions about recapturing the content, the EVA Project (Lounamaa and Salonharju 1999) at the University of Helsinki considers the burden on its system resources and the burden of its robots on the sites from which the content would be recaptured.

The National Library of Medicine's Permanence Rating System (see Section 2.1.1) also provides information to support limited recapturing.

2.2.2 Gathering Procedures

There are two general ways in which the archive acquires material. The producer can submit the material to be archived, or the archive can gather the material proactively.

In the first method, the best practices identified in the earlier section on creation become extremely important. Even within an organization, where the producer and the archive are almost one and the same organization, attention to standardization and limitations on the number of formats will have a significant impact on the ease with which submissions can be processed.

In the second approach, the archive may or may not have a formal relationship with the creator or the producer. In this gathering approach, the information to be archived is hand-selected or harvested automatically. In the case of the NLA, sites are identified, reviewed, hand-selected, and monitored for their persistence before being captured for the archive.

In contrast, the Royal Library, the National Library of Sweden, automatically acquires material by running a robot to capture sites for its Kulturaw3 project (Royal Library n.d.). The harvester automatically captures sites from the .se country domain and from foreign sites with material about Sweden, such as travel information or translations of Swedish literature. While the acquisition is automatic, priority is given to periodicals, static documents, and HTML pages. Conferences, usenet groups, ftp archives, and databases are considered lower priority.

2.3 Data Management: Metadata for Preservation

Metadata is needed to preserve the object and for users in the future to find and access it. Metadata supports organization, preservation and long-term access. This section deals with metadata for preservation. Other issues surrounding metadata for description and discovery were covered in the previous lecture on Metadata for Electronic Information Resources.

Archiving and preservation require special metadata elements to track the lineage of a digital object (where it came from and how it has changed over time), to detail its physical characteristics, and to document its behavior in order to reproduce it on future technologies. Each of the major preservation projects – Cedars, PANDORA, NEDLIB, the Harvard Library Project, etc., had its own set of metadata that it considered important for preservation. In 2001, the Research Libraries Group and OCLC reviewed the various sets of preservation metadata and concluded that there was sufficient similarity among the elements that a core set of metadata for preservation could be identified (OCLC/RLG 2001).

In 2001-2002, the Preservation Metadata Working Group developed a draft set of over 20 elements and numerous sub-elements for metadata preservation in the framework of the OAIS Reference Model (OCLC/RLG 2002). OCLC is already using the set as the basis for its Digital Archive and for the work that has been done with the U.S. Government Printing Office. In order to gain consensus on this set and to provide operational and implementation guidance, a follow on group, PREMIS, the PREservation Metadata: Implementation Strategies working group was formed (OCLC 2004b). Two subgroups have been formed, one on the elements and the other on implementation. The implementation group recently concluded a survey of operational preservation systems. The draft element set for preservation metadata and the results of the implementation survey are expected to be published in late 2004.

2.4 Archival Storage: Formats for Preservation

A major issue for the archiving community is which format(s) should be used for archival storage. Should the electronic resource be transformed into a format more conducive to archiving? Is the complexity of an interactive journal necessary or should it be simplified? Should the organization create a dark archive of archival copies in one format and a light archive for dissemination, which might be in a different format? Should consideration be given to the re-use of information and its enhancement or representation in more advanced access technologies in the future? Should the goal be complete replication of the electronic resource or should preservation provide a copy that is "just good enough"? (For example, Cedars has identified the concept of "significant properties," which are properties that are absolutely required in order for a user in the future to get the information value from the resource (Russell 2000).)

Of course the answers to these questions differ by resource type, and there is little standardization at this point. Most electronic journals, reference books, or reports use TIFF image files, PDF, or HTML. TIFF is the most prevalent for those organizations that are involved with conversion of paper issues of journals. For example, JSTOR, a non-profit organization that supports both storage of current journal issues in electronic format and conversion of back issues, processes everything from paper into TIFF and then scans the TIFF image. The OCR, because it cannot achieve 100% accuracy, is used only for searching; the TIFF image is the actual delivery format that the user sees. However, this does not allow embedded references to be active hyperlinks.

SGML (Standard Generalized Mark-up Language) is used by many large publishers after years of converting publication systems from proprietary formats to SGML. The American Astronomical Society (AAS) has a richly encoded SGML format that is used as the archival format from which numerous other formats, including HTML and PDF, are made (Boyce 1997).

For purely electronic documents, Adobe's PDF (Portable Document Format) is the most prevalent format. This provides a replica of the Postscript format of the document, but relies upon proprietary encoding technologies. PDF is used both for formal publications and grey literature. While PDF is increasingly accepted, concerns remain for long-term preservation and it may not be accepted as a legal depository format, because it is a proprietary format. Therefore, Adobe, the Association for Information and Image Management (AIIM) and several other organizations have developed a draft standard for archival PDF, called PDF-A. This provides a file specification for a minimal set of PDF features and functions that will continue to be migrated from one version of PDF to another. The draft is currently in the ISO process.

Preserving the "look and feel" is difficult in the text environment, but it is even more difficult in the multimedia environment, where there is a tightly coupled interplay between software, hardware and content. The University of California at San Diego (UCSD) has developed a model for object-based archiving that allows various levels and types of metadata with separate storage of the multimedia components in systems that are best suited to the component's data type. The UCSD work is funded by the U.S. National Archives and Records Administration and the U.S. Patent and Trademark Office.

2.5 Preservation Planning: Migration and Emulation

Preservation planning is the bridge between the decisions made about archival storage of the bits and bytes and issues of future access and user needs. There is no common agreement on the definition of long-term preservation, but some have defined it as being long enough to be concerned about changes in technology and changes in the user community. This may be as short as 2-10 years.

Two strategies for preservation are migration and emulation. Migration means copying the object to be archived and moving it to newer hardware and software as the technology changes. Migration is, of course, a more viable option if the organization is dealing with well-established commercial software such as Oracle or Microsoft Word. However, even in these cases migration is not guaranteed to work for all data types, and it becomes particularly unreliable if the information product has used sophisticated software features. Unfortunately, this level of standardization and ease of migration is not as readily available among technologies used in fields of study where specialized systems and instruments are used.

Emulation, a strategy that replicates the behavior of old hardware and software on new hardware and software, is being considered as an alternative to migration. There are several types of emulation. Encapsulation would store information about the behavior of the hardware/software with the object. For example, a MS Word 2000 document would be labeled as such and then metadata information would be stored with the object to indicate how to reconstruct the document at the engineering – bits and bytes – level. An alternative to encapsulating the behavior with every instance is to create an emulation registry that uniquely identifies the hardware and software environments and provides information on how to recreate the environment. Each instance would point to the registry (Rothenberg 1999; 2000). Taking emulation a step further is the idea of creating a virtual machine – a new machine that based on the information in the registry could replicate the behavior of the hardware/software of the past (Lorie 2001).

While the best practice for the foreseeable future continues to be migration, machine emulation has been tested with some success by the CAMiLEON Project, a joint project between the University of Michigan and the University of Leeds. However, Granger concludes that a variety of preservation strategies and technologies should be available. Some simple objects may benefit from migration, while others that are more complex may require emulation (Granger 2000; see also Holdsworth and Wheatley 2001).

2.6 Access

The life cycle functions discussed so far are performed for the purpose of ensuring continuous access to the material in the archive. Successful practices must consider changes to access mechanisms, as well as rights management and security requirements over the long term.

2.6.1 Access Mechanisms

The way in which access is viewed depends on the purpose of the archive, the audiences it will serve and the anticipated needs of those audiences over the long term. For example, national and institutional archives must be concerned with the ability to provide long-term access to the electronic information in a way that virtually replicates the look and behavior of the object today. This is a requirement because of the legal functions served by these archives of record.

Other organizations are interested in how they might actually improve access to current information in the future. A major reason for storing the information related to the U.S. National Library of Medicine's Profiles of Science materials in TIFF and other standardized forms, such as tagged ASCII, is so that the information can be re-purposed or enhanced. Even in its development stage, the project was able to improve the quality of the video clips by converting them to High Definition Video. The belief is that there will always be newer and better technologies, and a goal of the archive is to be able to take advantage of these advances in the future.

2.6.2 Rights Management and Security Requirements

One of the most difficult access issues for digital archiving involves rights management. What rights does the archive have? What rights do various user groups have? What rights has the owner retained? How will the access mechanism interact with the archive's metadata to ensure that these rights are managed properly? How will access rights be updated as the copyright status or security level of the material changes? Numerous groups, including the IEEE, ContentGuard, and MPEG, are developing digital rights management standards including expression languages to support interoperability in e-commerce transactions.

3.0 EMERGING STAKEHOLDER ROLES

A number of stakeholders can be identified including creators/authors, publishers, libraries, archives, Internet service providers, secondary publishers, aggregators, and, of course, users (Haynes, et al. 1997; Carroll & Hodge 1999; Hodge 2000; Hodge & Frangakis 2004). The roles these various stakeholders will play in the archiving process described above remains unclear, but there are several types of electronic information for which some patterns of responsibility are emerging.

In the early stages of the digital age, most electronic journal publishers considered the creation of an electronic archive to be the same as the internal production system. However, many publishers have since come to realize that archiving and production are not one and the same function. In some cases, they are quite antithetical.

The current environment shows a growing understanding of the need for archiving and long-term preservation among the major electronic journal publishers. This may not be the situation with smaller learned society publishers, but that may be more an issue of economics than of desire. The major electronic journal publishers such as Elsevier, Nature and Blackwell have committed to preservation using national libraries or trusted third parties.

Librarians and archivists, particularly those at national libraries, were early advocates of digital preservation issues. Many national libraries spearheaded initiatives and research projects without additional funds and without legislative mandates to cover digital deposit. In most cases, these projects have been instrumental in advancing the research and implementation of operational systems. For example, the Dutch National Library serves as the archive for Elsevier journals.

In addition to new roles for publishers and librarians/archivists, trusted third party archives are emerging. These third parties, such as the OCLC Digital Archive (2004a), JSTOR (2004), PubMed Central (2004), and BioMedCentral see archiving/preservation as an additional business/service opportunity.

A significant outgrowth of the OAIS Reference Model has been RLG's development of attributes of an OAIS-compliant archive (RLG 2001). This check-list can serve to assure a library, publisher or other organization that a particular third-party archive meets minimal requirements for import/export and basic functionality related to the other aspects of an archive.

Another significant development in the emergence of clearer stakeholder responsibilities, particularly for commercially published materials, is a January 2002 announcement on digital preservation by the International Federation of Library Associations and Institutions (IFLA) and the International Publishers Association (IPA). The draft presented for discussion highlights the importance of "born digital" materials and suggests that the appropriate place for preservation of last resort is with the national libraries. It is hoped that

additional legislative/policy efforts and funding for cooperative initiatives will result from this statement and from the inclusion of digital preservation on the agendas of these two major international stakeholder organizations.

4.0 SYSTEMS DEVELOPMENT

Since early in the investigation of digital preservation, institutions concerned about preservation and interested in performing this function have been awaiting "off the shelf" systems or services that could be installed with limited resources but variant levels of flexibility to meet local needs. These systems are beginning to become available from a variety of organizations. Several of the highlighted systems have or are developing "turn-key" or generalized systems that can be implemented by others. These are available both commercially and as open source software.

4.1 DSpace Institutional Digital Repository System

The DSpace Institutional Digital Repository System began as a joint project of the MIT Libraries and Hewlett-Packard Co. The architecture for the system is based on a number of preceding projects including those at Cornell, CERN, OCLC, LC and OAIS. DSpace 1.1 was released in November 2003 via an open source license (available from SourceForge). The MIT Libraries' implementation of DSpace defines various levels of support for different input formats. For example, "Supported" means that the format is recognized and the institution is confident that it can make the format useable in the future through whatever technique is desirable (emulation, migration, etc.). Note that there is no attempt to dictate the preservation method. "Known" means that the format is recognized and the institution will preserve the bitstream as-is, without a complete guarantee that it will be able to render the object completely in the long-term future.

In addition to these components of DSpace that are specifically preservation oriented, the DSpace suite includes search and browse capabilities and support for the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). This allows DSpace sites to harvest metadata from several sources and to offer services based on the metadata that is harvested.

4.2 Digital Information Archive System

The Digital Information Archive System (DIAS) is a commercially available system, originally developed to handle the electronic deposit of electronic documents and multimedia files for the Koninklijke Bibliotheek (KB) (the National Library of the Netherlands) (IBM, 2003a). It is based on the results of the various NEDLIB Projects led by the KB over the last several years. In the current DIAS system, IBM addressed the initial ingest, transformation, storage and metadata creation. The technical issues related to long-term access are being studied by IBM and are not a part of the December 2002 implementation. The DIAS system was implemented as KB's Deposit of Netherlands Electronic Publications (DNEP) system in December 2002, making it the first system of its kind (IBM, 2003b). KB's initial implementation is for e-journal publishers to deposit e-journals, but the plan is to extend this to other types of e-materials such as e-books.

In May 2003, the KB announced that it had signed an agreement with Kluwer to archive electronic journals featured on Kluwer Online Web Site. As of May, this contained 235,000 articles from 670 journals. The collection from Kluwer is expected to grow by more than 70,000 additional articles. The KB is seeking to enter into similar agreements with other publishers. Currently, the users (members of the public) must access the system from within the library because of copyright issues.

In 2003 the KB started a joint project with IBM to develop the preservation subsystem of DIAS. The work began with a series of studies around key preservation issues such as authenticity, media migration management, archiving of web publication, and a proof of concept of the Universal Virtual Computer. This subsystem will consist of a preservation manager, a preservation processor and tool(s) for permanent access. The Preservation Manager will manage and control the long-term durability of the digital objects using technical metadata. This is considered to be an essential part of the DIAS solution, since technical metadata will allow a future hardware environment to take the software bit stream and the content bit stream and provide access to the content. The problem that remains to be addressed is the obsolescence of the hardware of the rendering environment. Two major approaches are emulation and the use of a basic virtual computer. The aim is to have the turnkey system able to be generalized to other libraries and archives. Therefore, the system must be independent of the either of these preservation strategies.

4.3 OCLC Digital Archive

As an outgrowth of the preservation services that OCLC has provided to its member libraries for many years, OCLC has developed the OCLC Digital Archive. It provides long-term access, storage and preservation for digital materials, or "objects." The system is based on the OAIS. Records can also be ingested in batch. Currently the OCLC Digital Archive can ingest text and still images in formats such as PDF, HTML, TEXT, JPEG, BMP, GIF and TIFF. The goal is to accept more input formats in the future. This system is connected to OCLC's Connexion cataloging system, and the cataloger begins by creating a WorldCat record for the object, followed by a record that includes the preservation metadata. The preservation metadata is based on the early RLG/OCLC work in this area. These two records are linked (OCLC 2004a). In principle, it follows the Metadata Encoding and Transmission Standard (METS) structure, providing for descriptive, administrative, technical and structural metadata.

The system also includes an Administration Module that allows the user to modify existing records. The Administrator can also set privileges for a variety of functions so that various pieces of the metadata creation, ingest and dissemination processes can be assigned to different people with proper security. The Administration Module also allows the administrator to create collections and user groups for specific end-user access to the metadata and the content. Virus and fixity checks are run and results are reported through both the Administration and the cataloging (Connexion) modules.

4.4 PANDORA Digital Archiving System (PANDAS)

The PANDAS (PANDORA Digital Archiving System) has been operational since June 2001 (NLA 2003). The second version was installed in August 2002. Prior to the development of its own system, PANDORA tried to buy an archiving management system. From the response to the Request for Information, it became apparent that there was no affordable system on the market that met the requirements and so NLA decided to build the system in-house.

PANDAS enabled PANDORA to increase the efficiency of capturing and maintaining the archived Australian online publications and therefore, PANDORA's productivity. It also provides PANDORA's partners, primarily the state libraries, with more effective Web-based software for contributing to PANDORA.

The NLA has received a number of requests for access to the PANDAS software, since the current software options to support the creation and management of digital archives are limited. UKOLN recommended use of PANDAS for pilot web archiving projects it proposed for both Wellcome Trust and JISC (Day 2003). In response, PANDORA will soon make available an evaluation module, which will allow interested parties to have trial access to PANDAS.

4.5 Lots of Copies Keep Stuff Safe (LOCKSS)

LOCKSS (Lots of Copies Keep Stuff Safe) is an automated, decentralized preservation system developed by Stanford University to protect libraries against loss of access to digital materials (LOCKSS n.d.). LOCKSS development is supported by the National Science Foundation, Sun Microsystems, and the Mellon Foundation. LOCKSS software, which is free and open-source, is designed to run as an "Internet appliance" on inexpensive hardware and to require minimal technical administration. LOCKSS has been operational at Stanford for five years and the production version of the software was released in April 2004.

LOCKSS creates low-cost, persistent digital "caches" of authoritative versions of http-delivered e-journal content at institutions that subscribe to that content. LOCKSS uses the caching technology of the web to collect pages of journals as they are published, but unlike normal caches, the cached pages are never flushed. The LOCKSS server runs an enhanced web cache that collects new issues of the e-journal and continually compares its contents with other caches via a peer-to-peer polling system. If damage or corruption is detected in an institution's cache it can be repaired from the publisher or from another cache. LOCKSS safeguards the institution's access to the content while enforcing the publisher's access control systems and, the LOCKSS model generally does not harm the publisher's business model since it is based on the original subscription to the e-journal.

LOCKSS is moving toward becoming a self-sustaining alliance. "The LOCKSS Alliance will provide a small core of central support for technology, collections, and community services. In addition to a range of specific services, the Alliance will transfer knowledge, skills and responsibility for the LOCKSS Program from Stanford University" (Hodge & Frangakis 2004).

4.6 Fedora™ (Flexible Extensible Digital Object Repository Architecture)

The University of Virginia Library has teamed with Cornell University's Digital Library Group to develop Fedora, an open-source digital repository architecture on which a variety of digital library implementations can be based (University of Virginia Library 2003). Similar to DSpace, Fedora is focused currently on repository development and management. However, it will eventually include preservation services.

Fedora 1.0 was released as open source software (Mozilla Public License) in May 2003. Release 1.2 was made available in December 2003 (Johnston 2003). The first phase production repository based on Fedora will be launched in 2004. However, all the functionality described in the original design proposal will not be completed until 2005. The largest implementation of Fedora is at the University of Virginia Library's Central Digital Repository. A 2001 Mellon Foundation grant allowed for joint development of a production-quality system by Cornell and the University of Virginia. The system currently includes XML objects, text (full text and page images of e-books) and images in multiple resolution (Payette 2003). A number of other institutions and organizations are using or evaluating Fedora, including The British Library, the National Library of Portugal and the Thailand Office of Defense Resources. Fedora is a component of the DSpace architecture.

5.0 TRENDS AND ISSUES

The trend in archiving and preservation has moved from theoretical discussions to pragmatic projects and operational systems. There are more initiatives focused on the realistic details of metadata, selection criteria, technologies and systems for archiving. While the need to raise awareness has not completely disappeared, more time is being spent on partnership development, testing and implementation.

The focus of research and development has shifted to "filling in the gaps." The National Science Foundation (NSF) and the Library of Congress have announced research grants in areas such as extremely large data sets and long-term access to complex multimedia objects. The International Internet Preservation Consortium's Deep Web Working Group is investigating the capture of the dynamic web.

In addition to the trend toward pragmatic initiatives, cooperation has increased among projects and across stakeholder groups. OCLC, the UK's Digital Preservation Coalition (JISC) and RLG have been instrumental in identifying, supporting and advancing key areas of cooperation. As a real sign of maturity, the work is being "divided up". While some projects are developing operational systems, others are working in the background to achieve consensus on standards among/between projects. Unlike many standards activities in the past that have developed from local and regional practices, the work related to digital preservation is starting with the goal of international consensus.

Another key issue for electronic libraries is intellectual property rights. A recent study shows that Canada, Denmark, New Zealand, Norway, South Africa, and the United Kingdom have enacted legislation or have a legislative process in place that covers some form of digital publications (Hodge & Frangakis 2004).

Despite these positive trends, key issues remain. The cost of archiving and the lack of established business models that will sustain long-term preservation may prove to be significant stumbling blocks in the advancement of the cause of preservation. However, even these issues are being addressed in a pragmatic fashion. OCLC, Stanford University Libraries/HighWire Press, JSTOR, and major publishers such as Elsevier are actively dealing with questions of cost and how and who will pay for the archiving. Projects such as the archive of Elsevier material at Yale Library (also funded by the Mellon Foundation) (Hunter 2002) identified practices that can accommodate the access needs of libraries and users while meeting the economic requirements of producers. The development of value-added services that can help to subsidize basic archiving and preservation activities is being considered.

6.0 LOCAL INSTITUTIONAL RESPONSES

Many of the projects highlighted in this paper are national, regional or even global in scale. However, what can a local institution do to ensure the preservation of electronic resources?

First, it is important to be aware of what is going on in this field. What are the outcomes of the major projects? How are standards being developed?

There are several sources for this information. The major projects have extensive web sites, and many like Cedars and NEDLIB have produced numerous publications, which are available from the web sites. Secondly, the PADI site at the NLA (PADI, n.d.) and the joint DPC/PADI *What's New in Digital Preservation* site (Digital Preservation Coalition 2004) are major portals to digital preservation information. The Electronic Resources Preservation and Access Network (ERPANET 2004) provides workshops and reports on these workshops. Newsletters such as *RLG DigiNews* from the Research Libraries Group are excellent sources of up-to-date information about projects. The CODATA Working Group on Digital Data Preservation and the International Council for Scientific and Technical Information are developing a portal for resources related to the preservation of digital data sets.

The local librarian should take every opportunity to raise awareness about the importance of digital preservation at his or her institution. When possible, be proactive in seeking funds to start small projects for preserving digital materials. A concrete way to raise awareness is to ensure that archiving and preservation are

considered when negotiating licenses for electronic resources, such as electronic journals and databases. With many national regimes for deposit of digital materials lagging behind the practical uses of these materials, it is important to address these archiving issues in license agreements. Equally, it is important to try to establish a balance between the rights of the rights holders and those of the library and users.

The major lesson is to think globally but to act locally – scaling the findings of the major global activities to the local needs.

7.0 CONCLUSIONS

A review of the cutting-edge projects shows the beginning of a body of best practices for digital archiving. The early adopters in the area of digital archiving are providing lessons that can be adopted by others in the stakeholder communities. Through the collaborative efforts of the various stakeholder groups – creators, librarians, archivists, funding sources, and publishers – a new tradition of stewardship will be developed to ensure the preservation and continued access to our intellectual heritage.

8.0 REFERENCES

Beagrie, N. and Jones, M. (2001). *Preservation management of digital materials: a handbook*. London: The British Library.

Beck, J. (2003). PubMed Central and the NLM DTDs. *Presented at the ASIS&T DASER Summit, November 12-23, 2003, Cambridge, MA*. [Online]. Available: http://www.asis.org/Chapters/neasis/daser/Jeff_Beck_presentation.ppt [5 July 2004].

Boyce, P. (1997, November). Costs, archiving, and the publishing process in electronic STM journals. *Against the Grain*, *9*(5): 86. [Online]. Available: http://www.aas.org/~pboyce/epubs/atg98a-2.html [21 April 2004].

Byrnes, M. (2000). Assigning permanence levels to NLM's electronic publications. *Presented at Information Infrastructures for Digital Preservation: A One Day Workshop, Dec. 6, 2000, York, England.* [Online]. Available: http://www.rlg.org/events/pres-2000/infopapers.html/byrnes.html [21 April 2004].

CAMiLEON: Creating creative archiving at Michigan & Leeds: Emulating the old on the new. (2001). [Online]. Available: http://www.si.umich.edu/CAMILEON/ [21 April 2004].

CCSDS (Consultative Committee for Space Data Systems). (2001). Reference model for an Open Archival Information System (OAIS). Red Book CCSDS 650.0-R-2, June 2001. [Online]. Available: http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html [21 April 2004].

Cedars: CURL Exemplars in Digital Archives. (n.d.). [Online]. Available: http://www.leeds.ac.uk/cedars/ [21 April 2004].

Carroll, B. and Hodge, G. (1999). Digital electronic archiving: The state of the art, the state of the practice. [Online]. Available: http://www.icsti.org/conferences.html [21 April 2004].

Day, M. (2003). Collecting and preserving the World Wide Web: A feasibility study undertaken for the JISC and Wellcome Trust. [Online]. Available: http://library.wellcome.ac.uk/assets/WTL039229.pdf [21 June 2004].

Digital Preservation Coalition. (2002). [Online]. Available: http://www.dpconline.org/graphics/index.html [21 April 2004].

Digital Preservation Coalition. (2004) DPC/PADI What's new in digital preservation. [Online]. Available: http://www.dpconline.org/graphics/whatsnew/ [21 June 2004].

ERPANET: Electronic Resource Preservation and Access NETwork. (2004). [Online]. Available: http://www.erpanet.org [21 April 2004].

Granger, S. (2000). Emulation as a digital preservation strategy. *D-Lib Magazine,* 6(10). [Online]. Available: http://www.dlib.org/dlib/october00/granger/10granger.html [21 April 2004].

Haynes, D., Streatfield, D., Jowett, T. and Blake, M. (1997). Responsibility for digital archiving and long term access to digital data. JISC/NPO Studies on Preservation of Electronic Materials. [Online]. Available: http://www.ukoln.ac.uk/services/papers/bl/jisc-npo67/digital-preservation.html [4 June 2004].

Hodge, G. (2000). Digital archiving: Bringing stakeholders and issues together: A report on the ICSTI/ICSU Press Workshop on Digital Archiving. *ICSTI Forum 33.* [Online]. Available: http://www.icsti.org/forum/33/#Hodge [21 April 2004].

Hodge, G. and Frangakis, E. (2004). Digital preservation and permanent access to scientific information: The state of the practice. A joint report by the International Council for Scientific and Technical Information and CENDI. [Online]. Available with free registration: http://www.icsti.org/icsti_reports.html [21 April 2004].

Holdsworth, D. and Wheatley, P. (2001). Emulation, preservation and abstraction. *RLG DigiNews*, 5 (4), Feature #2. [Online]. Available: http://www.rlg.org/preserv/diginews/diginews5-4.html#feature2 [21 April 2004].

Hunter, K. (2002). Yale-Elsevier Mellon Project. [Online]. Availabile: http://www.niso.org/presentations/hunter-ppt_01_22_02/index.htm [21 April 2004].

IBM. (2003a). Digital Information Archiving System. [Online]. Available: http://www-5.ibm.com/nl/dias/ [21 June 2004].

IBM. (2003b). Royal Dutch Library preserves culture with Content Manager and DB2." [Online]. Available: http://www-5.ibm.com/nl/dias/resource/rdl.pdf [21 June 2004].

Inera Inc., (2001). E-journal archive DTD feasibility study. Prepared for the Harvard University Library, Office of Information Systems E-Journal Archiving Project. Pg. 62-63. [Online]. Available: http://www.diglib.org/preserve/hadtdfs.pdf [21 April 2004].

Internet Archive. (n.d.). [Online]. Available: http://www.archive.org [21 April 2004].

InterPARES: International Research on Permanent Authentic Records in Electronic Systems. (n.d.). [Online]. Available: http://www.interpares.org [21 April 2004].

JSTOR: The Scholarly Journal Archive. (2004). [Online]. Available: http://www.jstor.org [21 June 2004].

Johnston, L. (2003). Fedora™ and repository implementation at UVa. Presented at the DASER Summit, Cambridge, MA, 21-23 November 2003. [Online]. Available: http://www.lib.virginia.edu/digital/resndev/fedora_at_uva_DASER_files/frame.htm [21 June 2004].

LOCKSS. (n.d.) [Online]. Available: http://lockss.stanford.edu/index.html [21 June 2004].

Lorie, R. (2001, June). A project on preservation of digital data. *RLG DigiNews, 5* (3), Feature # 2. [Online]. Available: http://www.rlg.org/preserv/diginews/diginews5-3.html#1 [21 April 2004].

Lounamaa, K. and Salonharju, I. (1999, January). EVA-the acquisition and archiving of electronic network publications in Finland. *Tietolinja News, 1*. [Online]. Available: http://www.lib.helsinki.fi/tietolinja/0199/evaart.html [21 April 2004].

NEDLIB: Networked European Deposit Library. (2001). [Online]. Available: http://www.konbib.nl/nedlib [21 April 2004].

NLA (n.d.). Selection of online Australian publications intended for preservation by the National Library of Australia. [Online]. Available: http://pandora.nla.gov.au/selectionguidelines.html [21 April 2004].

NLA. (2003). PANDAS Manual. [Online]. Available: http://pandora.nla.gov.au/manual/pandas/index.html [21 June 2004].

National Library of Canada, Electronic Collections Coordinating Group. (1998). Networked Electronic Publications Policy and Guidelines. [Online]. Available: http://www.nlc-bnc.ca/9/8/index-e.html [21 April 2004].

OCLC. (2004a). OCLC Digital Archive. [Online]. Available: http://www.oclc.org/digitalarchive/default.htm [21 June 2004].

OCLC. (2004b). PREMIS (PREservation Metadata: Implementation Strategies. [Online]. Available: http://www.oclc.org/research/projects/pmwg/ [21 June 2004].

OCLC/RLG Working Group on Preservation Metadata. (2002). Preservation metadata and the OAIS information model: A framework to support the preservation of digital objects. [Online]. Available: http://www.oclc.org/research/projects/pmwg/pm_framework.pdf [21 June 2004].

OCLC/RLG Working Group on Preservation Metadata. (2001). Preservation metadata for digital objects: A review of the state of the art. [Online.] Available: http://www.oclc.org/research/pmwg/presmeta_wp.pdf [21 June 2004].

PADI: Preserving Access to Digital Information. (n.d.). [Online]. Available: http://www.nla.gov.au/padi/ [21 April 2004].

PANDORA. (n.d.) [Online]. Available: http://pandora.nla.gov/au/index.html [21 April 2004].

Payette, S. (2003). The Fedora Project. Presented at the DLF Forum, 17 November 2003. [Online]. Available: http://www.fedora.info/presentations/DLF-Nov2003.ppt [21 June 2004].

PubMed Central: a Free Archive of Life Science Journals. (2004). [Online]. Available: http://www.pubmedcentral.nih.gov/ [21 April 2004].

RLG. (2001). Attributes of a trusted digital repository for digital materials: Meeting the needs for research resources. [Online]. Available: http://www.rlg.org/longterm/attributes01.pdf [21 April 2004].

Rothenberg, J. (1999, January). Avoiding technological quicksand: Finding a viable technical foundation for digital preservation. Report to CLIR. [Online]. Available: http://www.clir.org/pubs/reports/rothenberg/contents.html [21 April 2004].

Rothenberg, J. (2000, April). An experiment in using emulation to preserve digital publications. NEDLIB Report Series; 1. [Online]. Available: http://www.kb.nl/coop/nedlib/results/NEDLIBemulation.pdf [21 April 2004].

Royal Library. National Library of Sweden. (n.d.) Kulturaw3 – Heritage Project: Long term preservation of published electronic documents. [Online]. Available: http://www.kb.se/ENG/kbstart.htm [21 April 2004].

Russell, K. (2000). Digital preservation and the Cedars Project experience. Presented at Preservation 2000: An International Conference on the Preservation and Long-Term Accessibility of Digital Materials, York, England, December 7-8, 2000. [Online]. Available: http://www.rlg.org/events/pres-2000/russell.html [21 April 2004].

University of Virginia Library. (2003). UVA Library Central Digital Repository. [Online]. Available: http://www.lib.virginia.edu/digital/resndev/repository.html [21 June 2004].

Weinberger, E. (2000). Toward collection management guidance. (Draft) [Online]. Available: http://www.leeds.ac.uk/cedars/colman/CIW02r.html [21 April 2004].

Electronic Information Management and Intellectual Property Rights

Graham P. Cornish
Consultant, Copyright Circle
33 Mayfield Grove
Harrogate
North Yorkshire HG1 5HD
UK

Graham@copyrightcircle.co.uk

ABSTRACT

The paper examines the concept of "digital is different" or not as the case may be and how and if digital publications are different from their paper counterparts. The concept of copyright in a digital age is explored and various different interpretations of the concept and its application discussed. The linguistic problems of using words from a paper-based environment will be considered and such basic words as "copyright", "copy", "author", "publisher" and "user" are put in a new context. Ideas such as databases, fair use and exceptions are explored in their relationship to technological measures to control copyright material by owners. Technological devices to control access to copyright material are explained using examples from the CITED, COPYSMART, IMPRIMATUR, and COPYCAT projects of the European Union. Mechanisms such as fingerprinting, watermarking and stamping are compared from a user/owner point of view rather than as detailed technology. The impact of the latest EU directive on copyright and the information society is explained in detail and the complexities of implementing this directive in different legal regimes and cultural environments will be drawn out.

1.0 INTRODUCTION

Intellectual Property rights are often a difficult concept to grasp in a world which is intensely focused on the material world. Intellectual Property Rights (IPR) are not in themselves tangible objects and therefore are often overlooked, ignored or even dismissed by many working in areas where they are actually crucial to the exploitation of what is being made, invented or thought about.

2.0 THE BASICS

Essentially Intellectual Property is a concept to protect the creativity of the human mind and in most jurisdictions is divided into a series of different types. These will be examined briefly but the emphasis of this chapter will be on copyright. There are basically four elements in IPR although other, more subtle, divisions can be detected and refined in both law and practice. There is protection for invention, protection for manufacture, protection for design and protection for the expression of ideas. Many objects may contain more than one of these areas in them and will therefore enjoy multiple protection.

3.0 THE OWNERS' AND AUTHORS' RIGHTS

When we create something we do two things: we put something of ourselves into it and we become vulnerable to the outside world. Let's say you paint a picture and show it to someone. They may laugh at your efforts, criticize your brushwork or say it is inspirational. Whatever the reaction you will take it personally as either praise or criticism of yourself, not just your artwork. This is just as true of an internal

memo in the company or a scientific paper in a journal. You may have had a brilliant flash of insight into a real problem or simply made a fool of yourself by misunderstanding the company approach or the scientific evidence.

It follows that what we create we should also control. From this comes the idea of copyright because the action we want to control is *copying* our work, whatever form that copying may take. The need to control may be because we are worried that someone will alter our work in some way or because they may deprive us of some money. It is important to realize that the copyright is quite separate from the work in which it subsists. The fact that you buy a book does not give you any control over the copyright in the book. Let us once again consider the painting.

The painting an artist painted may be a good one and could sell for quite a lot of money. So the artist will want to control what happens to the actual painting and also what people do with it. Hanging it on the wall in the buyer's house is OK but what if they want to make postcards of it and sell them in the local gift shop? The artist will probably feel this is unfair and they are making money out of their ability to paint.

Therefore the law in most countries gives authors and owners a set of rights which vary from one type of IP to another. These can be summarized as follows:

Patent	Right to make the object patented
Design right	Right to prevent others using the design or making things from it
Trademark	Right to market goods or services under the label
Copyright	Right to copy, issue copies, perform, broadcast, translate or adapt.

In addition to these essentially "economic" rights many countries give authors certain "moral" rights. In the Anglo-Saxon legal tradition these are very weak as that system is essentially based on the economic value of IP, but in Roman law and related traditions authors enjoy certain "moral" rights as well. These co-called moral rights are essentially related to the integrity of the person creating the work and therefore, by extension, to the work itself. In brief these rights are:

(a) To have the author's name included when the work is published

(b) To prevent significant parts of the work being removed

(c) To prevent significant additions being made to the work

(d) To prevent significant alterations to the work

(e) To prevent someone else's name being added to the work

(f) To prevent works being attributed to someone when they did not create them.

These may seem rather philosophical and even ethereal in nature but they become central to many of the later discussions on managing these rights in an electronic environment.

3.1 Patents

A patent gives an inventor the right for a limited period to stop others from making, using or selling an invention without the permission of the inventor. Patents are generally interested in functional and technical aspects of products and processes, and must fulfill specific conditions to be granted. Most patents are for incremental improvements in known technology – evolution rather than revolution. The technology does not have to be complex. Patent rights are territorial; a UK patent, for example, does not give rights outside of the UK. Patent rights last for up to 20 years in many jurisdictions but terms can vary and renewal is often required in some circumstances. This is a quote from the UK government's Patent Office website at www.intellectual-property.gov.uk.

It is not always clear who owns an invention, and there are potential pitfalls with patents. Despite these, a patent can be of use to an inventor, and can also benefit other people. You can arrange, through means of a licence or sale to use another inventor's patent. Additionally, large amounts of information can be learnt from other people's patents.

Sometimes items which would seem clear candidates for registering a patent are not in fact registerable. For example, computer programs. It is possible to patent programs for computers which, when run on a computer produce a "technical effect". However, if a program does not produce a technical effect when run on a computer it is unlikely to be patentable. A technical effect is generally an improvement in technology, and needs to be in an area of technology which is patentable.

For instance, an improved program for translating between Japanese and English is not patentable because linguistics is a mental process, not a technical field. On the other hand a program which speeds up image enhancement may be patentable because it produces a technical improvement in a technical area.

3.2 Trade Marks

A trademark is any sign which can distinguish the goods and services of one trader from those of another. A sign includes words, logos, colours, slogans, three-dimensional shapes and sometimes sounds and gestures. A trademark is therefore a "badge" of trade origin. It is used as a marketing tool so that customers can recognize the product of a particular trader. To be registerable in many jurisdictions it must also be capable of being represented graphically, that is, in words and/or pictures. A trademark that is not exploited can lapse after a given period of time but if it is used and renewed if necessary it can last indefinitely.

3.3 Designs

Designs are protected in different ways in various countries but the EU has standard rules for protecting design right. A design refers to the appearance of the whole or a part of a product resulting from the features of, in particular, the lines, contours, colours, shape, texture and/or materials of the product itself and/or its ornamentation. Owners of design right may need to rely on copyright or industrial property rights of various kinds. Some countries offer protection by three legal rights;

- registered designs
- unregistered design right
- and artistic copyright

The design of a product can be synonymous with the branding and image of a company and can become an asset to them with a monetary value that could increase. Design registration usually gives the owner, a monopoly on his or her product, i.e., the right for a limited period to stop others from making, using or selling the product without their permission and is additional to any design right or copyright protection that may exist automatically in the design. Registering a design gives the owner the right to take legal action against others who might be infringing the design and to claim damages. Registering can also deter a potential infringement and also brings the exclusive right to make, import, sell or hire out any article to which the design has been applied or to let others use the design under the terms agreed with the registered owner.

3.4 Copyright

The idea behind copyright is rooted in certain fundamental ideas about creativity and possession. Basically, it springs from the idea that anything we create is an extension of 'self' and should be protected

from general use by anyone else. Coupled with this is the idea that the person creating something has exclusive rights over the thing created, partly for economic reasons but also because of this extension of 'self' idea. Copyright is therefore important to ensure the continued growth of writing, performing and creating. Copyright law aims to protect this growth but, at the same time, tries to ensure that some access to copyright works is allowed as well. Without this access creators would be starved of ideas and information to create more copyright material. Copyright is not something that can be registered: the Berne Convention for the Protection of Literary and Artistic Works, to which most countries belong, prohibits registration as a condition of claiming copyright. Copyright does not often last indefinitely. As a general rule it expires 70 years from the end of the year in which the author dies although there are different rules in many countries for works without authors, unpublished documents and those created by government or other institutions. From the point of view of electronic information copyright is by far the most important right. Table 1 summarizes the types and characteristics of intellectual property rights.

Table 1: Types and Characteristics of Intellectual Property Rights

Type of IP	*Protects*	*Lasts for*	*Registerable or not*
Patents	inventions	20 years approx.	registerable
Design right	appearance	10-15 years	can be registered or not
Trade marks	distinguishing sign	indefinite	registerable
Copyright	expressions of creativity	70 years after death	not registerable

Note: This table gives examples only and is not a legal guide to any jurisdiction.

3.4.1 Rights in an Electronic World

Apart from some possible uses of computer software, patents are not relevant to the electronic *information* world as they relate to manufacturing objects. Design right applies also to making things although the documents containing the design will almost certain be electronic in today's processes. Trademarks will be highly relevant in terms of branding services and products such as Dell Computers, AOL online or Netscape Navigator as well as the thousands of products about which information is available through the Web. The greatest complexity in IP terms for electronic information is copyright and this chapter will concentrate on this issue although the other rights will be discussed as appropriate where they arise.

It is important to realize that whatever is said about "electronic rights" is usually derived from fundamental principles discussed above and is often the result of a long and complex evolutionary process which has seen copyright, for example, changing from a right to protect publishers to one that protects authors (Feather 1994). Although supplementary laws have been passed in many countries to accommodate technological developments these invariably build on existing principles. The EU Directive on the Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society does try to introduce additional rights but these are still built on the basic ideas behind copyright as such (European Parliament 2001).

3.5 Databases

Some materials are not considered by all legislations to be suitable for copyright protection. For example, directories, lists of organizations or people or simple bibliographies. Nevertheless such items are the result of considerable investment in terms of money, labour and technical skill. For this reason the EU introduced a directive to regulate this situation. As a consequence any thing qualifying as a database is now protected in the EU by a special Database right. This lasts for only 15 years instead of the usual 70 but is capable of being automatically renewed or extended each time the database has significant changes made to it. Thus a dynamic database will be continually updated and therefore the 15 year protection "clock" will continue to tick until the time when the database is shut down and no longer active.

This right protects all kinds of valuable products which do not demonstrate any kind of creativity in terms of original thinking.

3.6 Ownership and Protection

Although these rights are intangible in themselves, they nevertheless exist in law and usually have the same status as any other property right. They therefore have to be protected from unlawful use and exploitation in the same way as any other property such as land, equipment or buildings. This means that IP is essentially a matter of law relating to personal property and is dealt with according to the legal tradition of the country concerned. Major acts of piracy will probably be dealt with by the police on behalf of the owner but other misuse will have to be dealt with through civil procedures of one kind or another.

It is important to note that authors and owners may, or may not, be the same person or institution. Authors create (write, compose, paint, carve, draw, perform, etc.) and owners own the rights mentioned above. Usually the author is the first owner of the rights which the law gives but this may not be true of works created as part of the author's employment. In this case the rights belong immediately to the employer and this is true in nearly all countries. However, most authors cannot exploit their works in any commercial sense without the assistance of some kind of publisher, whether it is a conventional commercial publishing house, a government agency, a research institution or a website host. In return for making the work widely available the organization doing this will almost certainly want some rights over the work in return. As all IP is a tradable commodity this can be done in the form of outright selling, licensing for a specific time-span or purpose or format. Authors can set all kinds of conditions on the sale or lease of their IP rights and usually do so. However, this process immediately makes clear the distinction between the rights that *authors* enjoy and those that *owners* have and why different people may have different rights in the same work.

3.6.1 What Needs Managing?

Given that copyright is both an economic and personal (moral) right, why is there a need to manage it in any particular way? Essentially this is because the very nature of the property protected by copyright (and other IPRs) is intrinsically different from that protected by the usual laws governing theft, trespass and even fraud. Authors and owners have two sets of rights to consider and protect which is rarely the case in the purely material world. The mechanics behind making available physical goods and IPRs are fundamentally different. Selling, renting or lending a physical object deprives the owner of the use of that object either permanently or on a time-limited basis under agreed terms. Usually this means that money has been paid for the acquisition or temporary use of the item. Someone may buy a car or rent it for a holiday. In either case the original owner parts with the car and the new owner/holidaymaker has the right to use it instead. An object has passed from owner to user: 1-1 = 0 (the owner) and 0+1 = 1 (the user). In the case of many IPRs, especially copyright, the mathematics are quite different. The information is provided (free or at a price) by the original owner or creator (author). It then passes to the user but, in this case, the owner still retains the original information but the user (buyer) also has it. A market research company produces a report on how much coffee is produced in Brazil each year. This information is read by researchers, coffee marketers and economists who may have paid for it or heard it on the TV or simply red it in a newspaper. The information remains with the original owner but it has been spread to a large number of others *as well.* The mathematics are 1-1 = 1 (the owner) and 0+1 = 1 (the user) – a quite different outcome. For this self-evident reason (but it is always worth stating the obvious when dealing with a complex issue) special laws relating to IP are necessary and therefore different management techniques are needed to exploit and protect it properly.

3.6.2 Moral Rights

Authors enjoy the moral rights mentioned earlier – basically integrity of the work and acknowledgment as the author. These rights in the past have been managed in a very elementary way. Authors enjoyed them as

an absolute right in countries such as France and Germany but did not enjoy them at all in the UK and Ireland until well into the 1980s and then only on a very restricted basis. In the UK, for example, such rights have to be asserted in writing to the publisher and apply only to monographs and films. Neither do they apply to works created as part of employment. If such rights were infringed then the author has the absolute right to take the publisher to court to put the matter right and obtain damages.

3.6.3 Economic Rights

More far-reaching and of much greater significance to the commercial world are the economic rights that owners enjoy. Owners are more likely to be companies, publishing houses, film companies or sound recording firms than individuals. Owners of copyright enjoy a series of rights which vary somewhat from one country to another but can be summarized as follows:

- Copy the work
- Make the work publicly available
- Perform, show, play or broadcast
- Adapt or translate
- Lend or rent the work.

These rights were traditionally managed through the courts because it was comparatively easy to spot when someone had copied a work, whether it was a scientific paper, musical disc, photograph or film. If copying was identified then appropriate action could be taken. However, as most copying was, and often still is, of a small nature (individual copies for personal use, small parts of a film for demonstrating a technique or personal viewing at home) owners rarely took action unless this was thought to be symptomatic of much more extensive copying. To cope with this situation some countries (notably the UK, Ireland and Scandinavian countries) developed a concept of "fair" use of copyright works. This made legitimate the occasional use of works for various (and varying) purposes provided it did not seriously harm the interests of the copyright owner. This is actually allowed in the Berne Convention (Article 9.2). Owners often regard this "concession" as a result of what is called market failure. In other words, the owners have failed to find a way of meeting the needs of users by allowing them to make limited copies in return for permission and/or a fee. Therefore the market has failed to meet a need and so the law steps in and fulfils that need anyway.

Making a work public was also easy to detect and this was most often in the form of pirate versions of books, videos or discs. In the same way performing a work in public, showing a film or broadcasting something was easy to trace and appropriate action could be taken to recover any lost revenue and prohibit this happening again without the owners consent. This is equally true of lending and rental.

Transforming a work was often less easy to detect as the transformation may have been so extensive that the original could not be easily identified. Whilst we may think of this in terms of books being changed into plays or films (in the latter case the film often bears little resemblance to the original story!!) this is equally a problem with maps, computer software and scientific articles which are frequently plagiarized by students of all levels.

The only real course of action for owners is to threaten some kind of legal action once the infringement has been discovered. A warning letter will be sufficient for small offences but it may be necessary to take the offending person to court which is expensive for both owner and infringer and the outcome is never totally certain.

3.6.4 Owners' Needs

It is clear that all these remedies are retrospective. If an infringement is detected then action is taken against the offending person or institution and recompense is sought and steps taken to prevent the

infringement continuing (if appropriate). Such action is often taken, not only because of the economic damage of the actions actually committed but as a warning to others that the owner will not tolerate infringement of rights. Whilst remedies are appropriate, most owners would prefer to take proactive action to prevent infringement and ensure a proper respect for their property and economic returns on it when possible.

3.7 Language in the Electronic Environment

A major problem when talking about intellectual property in the electronic environment is the language used. Essentially the words traditionally used in publishing and writing are still used when talking about digital products or the Web. This means that there can be misunderstandings and a few examples will highlight the problems.

Copy. Whilst we may understand that a copy is a reproduction in some form of the original (photocopy, scanned image or even hand drawn) what is meant by a copy when using a website for example? Many "copies" are made between the homepage and the end-user, often via Internet Services Providers (ISPs) and even on to the user's hard disk but do they all count as copies or merely transfers of digital data? If someone sends an identical email to twenty people all at the same time are they twenty copies or twenty simultaneous transmissions of the same message?

Journal. The concept of the scholarly journal is well established in academic and scientific circles but what is an electronic journal? The journal is usually a package which contains various items such as articles, letters to the editor, advertisements, announcements of meetings and even a contents page or index. Subscribers buy the package whether or not they want everything in it. But in an electronic context such packaging is not necessary in one sense as the user can pick and choose which elements may be wanted or discarded. The "brand name" of a journal however may be vital (see the remarks on peer review) so the concept needs to continue but not as it is today.

Document supply. By this is usually meant the transfer of a copy of a document from one place to another, usually between libraries initially but for the benefit of the end-user. When something is sent digitally what is transferred? The image is certainly made available in another place but is it necessarily actually supplied for retention. It may be a system to allow temporary access, not dissimilar to actually lending something.

3.8 Managing in the Electronic Context

But owners have other interests besides enforcing the legal rights they enjoy in each country. In the current technological climate both creators and users of intellectual property have certain basic needs which must be satisfied if they are to be assured their material can be safely released in electronic form. Although it would be easy to categorize the needs of owners as *protection* and those of users as *access*, this is a very simplistic approach and many more requirements need to be examined before any comprehensive system of electronic control can be put into place.

3.8.1 Owners' Needs

In the case of rights owners the need for protection should not be seen so much as *preventing* use of their material as *controlling* that use (Van Slype & Van Halm 1988). Basically a rights owner needs to be able to control:

a) copying from paper on to paper

b) copying from paper into electronic formats

c) copying from electronic formats to paper

d) copying from electronic format to electronic format including storage and transmission

e) multiple copying.

Clearly items (a) and (e) are issues in the paper environment and (b) is primarily an issue for producers of paper documents whilst (c) and (d) are entirely new issues to be faced.

The rights owner may also wish to be able to operate:

a) Differential pricing for sector, group, type of use. For example, different parts of a database may be priced differently for access depending on the value of the content. A senior researcher of academic could be given access to the whole database because such a person would have the necessary economic power but access for a student may be more limited because of the cost involved or even because the rights owner does not wish some material to be made readily available to students for reasons of sensitivity or security.

b) Differential pricing for individual elements within a product such as specific journal articles, elements in a directory or areas of a classification arrangement.

These are entirely new concepts in publishing which could not be achieved in a paper world. The best that could be achieved was differential pricing of a total product such as a journal title. Different subscription rates for institutions, libraries and individuals are common. An example of this is Haworth Press in the USA which has had three levels of pricing for its products for many years.

The rights owner also needs:

a) Protection against unauthorized use by groups or sectors who may not have purchased any or all of the rights to do the actions mentioned above.

b) Protection against unauthorized use of products. Not everyone wants their music used to promote motor cars or their writings used in political campaigns!

c) A system which provides data on use for marketing purposes. Information intermediaries, by penetrating previously untapped markets, will be able to provide considerable data on potential marketing for products and services.

d) A system which will ensure compensation for all actions over which the rights owner has exclusive control. Collection of royalties and fees, even if set at zero levels, needs to be achieved efficiently and effectively.

Of these needs only (d) has been partially achieved in the paper world usually by the blunt instrument of collection societies. In the electronic context these are new concepts which need to be managed in new ways.

3.8.2 The Intermediary's Needs

In the past owners (and authors) rarely distributed their material directly. Most, although not all, publishers use distribution mechanisms to make their products available. Distributors are best described as "information intermediaries". They do not create intellectual property themselves nor do they directly publish the expression of it. Their role is to act as an intermediary between the producer (publisher) and the users of the published information. Information intermediaries may act directly between these two elements of the information chain or may themselves deliver to other information intermediaries for onward delivery to end-users. Most intermediaries can be considered as distribution and fall into a series of categories, most of which are non-exclusive. Distributors can be divided broadly into the following segments, although some interplay between the different segments is inevitable and the distinction between the different roles is becoming blurred.

(i) Booksellers are primarily concerned with buying printed books, and sometimes journals and newspapers, in bulk and selling on to the end-user (the public). Their role is still primarily as distributors of printed material. Some booksellers act as information intermediaries to sell on to other intermediaries especially libraries. Booksellers are therefore of very limited importance in the electronic context.

(ii) Subscription agents also act as distribution agents for publishers but also play a vital role in the economics of library management as they collect together subscriptions from many different libraries for a range of different publishers without either having to worry about the complexities of matching one to the other (Renwick 1991). Subscription agents mostly deal with other information intermediaries such as libraries rather than end-users. Increasingly subscription agents are becoming involved in direct document delivery systems. This can give subscription agents a new role as collective managers for electronic products. However, this is not popular and is a model which has not developed significantly.

(iii) Database hosts. Just as the number of journals has increased to a point where publishers could not reasonably sell direct to their customers, so the multiplicity of databases has led over the years to the development of the database host. The host may, or may not, generate copyright material directly but is mainly concerned to provide a mechanism through which other intellectual-property creators can distribute their electronic data. Hosts usually negotiate fees and conditions between both originators and users, whether intermediaries or end-users. Although, initially, CD-ROM producers tended to distribute their own products, the rapid increase in the number of CD-ROM databases available has led to some database hosts or other distributors acting on behalf of CD-ROM producers in the same way as subscription agents.

(iv) Libraries are information intermediaries because they collect and store large quantities of published material which they make available to their readers (end-users) in a form which enables them to exploit such materials for a wide range of purposes. Traditionally, libraries have been *passive* information intermediaries, that is, they have collected and stored information and organised it in a meaningful way but have left their end-users to exploit it as they saw fit. Increasingly, libraries are becoming *active* information intermediaries, providing detailed and analytical guides to the literature, producing information bulletins, current awareness services and proactive document delivery systems based on profiles of individuals interests and needs, through SDI (Selective Dissemination of Information) services. In the present world economic climate libraries are becoming much more commercial and are beginning to exploit their collections for financial gain rather than simply for the benefit of their readers. In addition to conventional libraries there are several organizations throughout the world which exist almost exclusively to offer document delivery services either to individuals or to other libraries. These include the British Library Document Supply Centre in the UK and INIST in Nancy (France). Other major players include the Central Medical Library in Cologne and the Technical Information Library in Hannover. These organizations specialize in document delivery using conventional library collections and the rapid increase in their business has put great pressure on the publishing industry to develop appropriate mechanisms to allow this considerable business to continue and increase in response to worldwide demand whilst ensuring protection of publishers' and authors' rights.

In theory the provision of material in electronic form could mean the end of libraries. It is unlikely that this will not happen because users cannot have access to every source of supply and need guidance on what is the best and most appropriate source for their needs. As has happened in the case of databases there will always be a need for an intermediary although the role for that intermediary will change but not disappear. Nevertheless the role of the librarian will change from supplying information and documents to supplying packages of information. Information is big business, however it is defined. The desire, never mind the need, for information is a constant feature

of current cultural patterns, particularly in the industrialized world. The information may be supplied in various ways: newspapers, journals and books, broadcasts, television, teletext or online system. Nevertheless the demand for it is there and document supply is merely one aspect of the way that demand is being satisfied. In the more sophisticated reaches of the information supply industry librarians are not simply renamed "information scientists" but transmogrified into "knowledge scientists". A knowledge scientist is not expected to provide information but to interpret it for the customer. This particular trend leads those in this situation to receive requests for appropriate data, suitably packaged, on a given topic or aspect of a topic. The resulting package may be a concoction of statistics, manipulated data, law, company information, economic projections and predictions and some documents. The knowledge scientist will be required to obtain such documents either locally or from remote sources. The customer in this situation has little interest in where or how the document was procured so long as it supplies the needs of the time. Although this is at one extreme end of the information supply spectrum, it is nevertheless symptomatic of an increasing trend in the information industry at all levels. Documents are increasingly seen as vehicles for information in its widest sense. And "information" should not be understood in too narrow a sense. The content of a well-established piece of non-factual writing (the distinction between fiction, novel and literature is an issue to be explored on another occasion!) is viewed by many as a piece of information and the format in which it is delivered is far less relevant than the delivery itself.

Therefore the role of the library in particular will be revolutionized in the electronic world but the concept of the library as "neutral" will remain. The role of the library is essentially that of a "neutral" intermediary between creators, owners and users of intellectual property. This possible role will be discussed later.

(v) Information brokers are usually commercial enterprises which identify the information needs of individuals and institutions and attempt to meet these needs through a range of services including tailored information packages, alerting services, information services tailored to specific needs and document delivery. Such organizations do not, as a rule, have collections of literature themselves but rely on libraries or other document suppliers for the individual items they have themselves identified for their clients (end-users). As their role is both to exploit for commercial gain and exploit materials for which they have not themselves paid any contribution to the publisher/author, this area is a primary one where management of IPRs is essential and highly beneficial to the owner.

Intermediaries have certain basic needs to fulfill their role in the information chain. These can be summarized in the electronic context as:

Information intermediaries need to be able to:

a) Gain access to a work

b) Store a work

c) Retransmit a work repeatedly in different formats depending on the needs of customer

d) Exploit additional markets to which rights owners may have no access

e) Provide additional services which publishers are unable or unwilling to develop

f) Protect any privileges they enjoy under national legislation to allow them to provide services to their clients. This is especially true of non-commercial libraries.

Clearly some of these conflict with the interests of the copyright owner unless they are carried out in a spirit of cooperation as discussed later. Of the items on this list (a) and (b) only are really relevant to the paper context where they are the normal part of library and information provision. Other items would need permission from the owner.

3.8.3 The User's Needs

The users of information, often called "end-users" to distinguish them from intermediaries who also use information but not for their own individual needs, have requirements for which copyright management must also cater. As already stated many legislations cater for the needs of users in a limited way by recognizing that payment or permission cannot be sought in every case. Therefore something has to be done to regularize this situation. Laws which say "you shall not…." but is incapable of enforcing that prohibition is bad law. End-users are the primary reason that any document is published. Publishing broadly means making available to the public and it is the reading public to which most publishing is aimed. End-users clearly need or want published material for a range of purposes including leisure, general information for daily life, education at all levels, intellectual research and industrial or commercial exploitation.

Until a few years ago, few end-users have had direct access to large quantities of published material except through an intermediary such as a library, but with the increasing use of online publishing this pattern is changing rapidly and end-users can now obtain access to electronic databases through terminals in libraries, research departments or directly in their homes.

End-users' needs can be identified as

a) Consult the work
b) Store the work
c) Be confident of the confidentiality of use activities
d) Be assured of the origin, originality and integrity of the document supplied
e) Ensure any privileges they enjoy under national legislation are protected

Unlike other players in the information chain, all the needs of users can be met in the paper world and, in theory at least, all of them could be under threat in the electronic world. Therefore they all need careful attention in any management exercise for electronic information.

3.8.4 Users' Privileges

As explained earlier market failure leads to a situation where many countries grant users of copyright material certain privileges to use the works in certain limited way without reference to the copyright owner who has no real right to object to these exceptions to their rights. These exceptions vary from one country to another and the recent EYU Directive on copyright has done nothing to harmonize these. Each EU member state may choose from a wide range of exceptions which can be implemented nationally if they wish. Such exceptions may apply to individuals for research, private study, personal interest or use, criticism, essay writing, educational purposes or study. Institutions such as libraries or schools may also be allowed to use material in limited ways for the benefit of their readers or pupils. Similarly the organs of government such as Parliament or departments may be allowed certain exceptions and it is usual that copyright cannot inhibit either justice, democracy or national security.

3.9 Some Possible Solutions

The question is whether such a system is even remotely possible or whether such a system is more like a dream than reality. However it has been said that "when one man dreams, it is a dream; when several men dream the same thing it is the beginning of reality." (Cornish and Keates, 1993).

The CITED "solution"

Fortunately several men and women had the same dream and joined together to form a consortium which they called CITED (Copyright in Transmitted Electronic Documents). The group formed a consortium which

applied to the European Commission under the 6th. Call for their ESPRIT II Programme under a Workpackage entitled "Electronic Copyright" which was accepted. The partners included electronic publishers, a computer manufacturer, a library, a lawyer, security and software specialists and experts in databases and networking.

The basic philosophy of the CITED project was that, since we are dealing with information which is stored and, more particularly, processed digitally, it is therefore possible, in the digital environment, to control the processes which are an inevitable part of digital technology and, in consequence, control the copying of copyright material. In the electronic context it is immaterial what information is represented by the digital signal in any given case; what was proposed was the development of a generic model of copyright protection of digital information (the CITED model) together with corresponding guidelines and toolkit to enable the model to be implemented in specified domains. The generic nature of the CITED model means that it can be relatively easily mapped on to the legal background both as it is currently and within its foreseen developments. The generic model is also capable of being used as a standard against which different systems can be tested to ensure that they confirm to the basic requirements of a CITED protected system.

The idea behind the project was that compliance with the model CITED could be established for a range of standards, and via a number of different technical strategies. The level of protection could be defined, depending on the nature of the information to be protected, and the rights of various CITED users could be specified. These rights are specific to the users, but the effective right to gain access to a particular piece of information can be made, in practice, to depend on the protection level of that information. At the technical level the CITED model is primarily concerned with the relationship between "actions" (i.e. those actions which users may wish to undertake), called in CITED jargon "events" and "rights" (i.e., those legal rights which owners, distributors and end-users enjoy. The CITED model is concerned to capture on record the "actions" and the response to these actions can then depend on the rights which users of the appropriate information have acquired by purchase or agreement. Although the primary method of acquiring rights is to purchase these, there is no reason why a CITED facility should not permit free access if the owner so wishes. What CITED would permit is the monitoring of the free use which would in itself be a valuable piece of data. Naturally a critical area is the detection of actions which are not permitted, either generally or to a particular user. Of course, CITED could be used both as a countermeasure to such threats and as a marketing tool. Attempts at unauthorized use could then bring not just a negative response but information as to how the action which has been refused could be executed. As described, the CITED environment could be therefore dynamic and could respond to a range of possibilities as described on the Project's website at http://www.newcastle.research.ec.org/esp-syn/text/5469.html.

Naturally some of the technical tools used in the CITED project are adopted from the repertoire developed for the security industry. However, within the CITED project these are viewed as placing a protective guard around the copyright information in a manner which, while preventing unauthorized copying, nevertheless permits convenient access for authorized use. In fact CITED is a sort of tool kit which provides a variety of implements which may be needed in some, but not all, environments.

The CITED project was never implemented in its entirety. The original concept of designing both software and hardware that would achieve all its goals proved impossible. However, it is important from the ERMS point of view because it set out the requirements of any system for managing electronic information and identified the discrete elements which any copyright owner or user could then consider independently for implementation. Given its theoretical nature, CITED could provide the model to handle a number of the issues identified above. When the project was first proposed in 1989 it was considered too futuristic and was the object of some mirth in the information management world. It is interesting to note how many of the CITED concepts have now come to fruition in DIFFERENT mechanisms!

3.9.1 Copying from Paper to Paper

This is not an action relevant to CITED protection mechanisms. It is, however, still one of the most common ways in which copyright material is copied, despite the great advances in electronic technology.

CITED solution: none.

Although this is not an issue as such in electronic information management several attempts have been made to design a paper which can be used for printing but which has chemical properties that will not allow photocopying. An even more elaborate scheme was to devise a font with a minute change made in a very common word (for example the word "the" in English which appears in nearly every sentence) and the photocopy machine or scanner would recognize this tiny change and produce an unreadable copy.

Neither of these solutions has found favour. In the first case the quality of the original material was impaired and in the latter the sophisticated technology and its installation into a large number of photocopy machines militated against the project in terms of cost when considered against (a) the volume of copying likely to be prevented and (b) the value of lost revenue to the publishers.

3.9.2 Copying from Paper into Electronic Formats

There is a rapidly increasing requirement for libraries, archives and private individuals to be able to digitize existing paper text which can then be stored and, on occasions, transmitted as required to users of the service or to other researchers and colleagues. This requirement is not universal but long experience shows that 80% of requests to libraries can be satisfied by 20% of a repertoire of journals. Therefore the need is not to digitize all journals, or even all articles in a journal, but to be able to digitize those that are in most demand. Naturally these are the titles that generate most income for publishers and a way to meet the needs of both parties is essential.

CITED solution: Copying from paper into electronic form could not be prevented using CITED mechanisms but such copying could be done with the agreement of rights owners. This agreement would be to install CITED protected mechanisms in the digitally-stored text to carry out CITED monitored activities such as further copying and distribution. Given the huge increase in digitization programmes in many parts of the world this issue is a serious one that needs addressing. Unfortunately many publishers and owners of rights are nervous about giving the initial permission in case the subsequent technology fails to manage and protect their interests adequately. Many digitization programmes are therefore limited to older material that is out of copyright or material where the rights are owned by the institution doing the digitization such as unpublished archival material or internal papers and reports. This material is often in low demand and the economics of digitization are against taking the programme forward. Projects such as JSTOR of which more details can be found on the web at http://uk.jstor.org/about/need.html which aims to digitize scholarly journals are limited to non-current journal issues which have lost much of their economic benefit to the publishers.

3.9.3 Copying from Electronic Formats onto Paper

This is a requirement for many all users of electronic information of all kinds and includes most libraries and other information intermediaries, as well as individual researchers. Although the temptation is always to think of online supply, the use of CD-ROM and electronically-based text frequently generates requirements for paper copies which can be used in many different environments. Although this can be done already under licence within an institution there are growing problems where documents are requested from a distance and the requester is unaware that the document required is in electronic form This problem is compounded when documents are published in both paper and electronic form which could lead to the anomaly that a paper copy could be made from another paper copy but not from the identical electronic text.

CITED solution: Copying from electronic format onto paper can be controlled through software already but this procedure needs to be capable of control, charging and monitoring depending on the requirements of the

copyright owner and the end-user's status. This is a primary example of how CITED protected text can achieve these goals. CITED mechanisms enable individual users or institutions such as libraries to copy onto paper in return for appropriate royalties. By requiring the user to have "bought in" to the system in advance the CITED model set the scene for many contemporary systems that enable different rates to be charged to different users. Also the same user can be charged different amounts for copying different articles. At the same time it is possible to permit copying by one group of users for low or no cost (students, for example) whilst charging a higher rate for the same activity to researchers in commercial institutions.

3.9.4 Downloading from One Electronic Format to Another

As more and more documents are available in electronic form, or only in electronic form, copyright owners and document suppliers will be able to deliver to end-users only by using the electronic forms available. Although copying from electronic formats onto paper is already controlled and well used, there is inevitably a need to be able to download into the user's own system.

CITED solution: Documents protected using the CITED model can have the facility to download or not built into their protection mechanism to permit copying in the same way as copying onto paper (already described) However, the model did not offer protection against further copying and distribution once the work has been downloaded.

These facilities were further developed in projects such as COPICAT and the highly sophisticated model described in Project IMPRIMATUR (Cornish 2000).

3.9.5 Meeting the Needs of the Information Supply Industry

The information supply industry is one which is growing in both size and complexity. The roles of different players are becoming less and less clear. Nevertheless the basic requirements described earlier remain the same. The CITED model offered the possibility of easy access, flexibility, comprehensive data and recompense for the owners of the many different rights involved in its operation. Although collection of royalties and data can be achieved through this system, it is desirable that such collection should not be done by each rights owner separately but could be achieved through a central agency similar to a Reproduction Rights Organization (RRO). It is still hoped that there may be ways to establish a Trusted Third Party (TTP) which would oversee and manage such elements of the system as lend themselves to centralization. there are still many areas of this model and its application to develop and the management of the concept is itself one of these.

3.10 Other Applications

Clearly CITED is capable of being applied to many different areas of information work. It can be applied in any digital environment such as sound recordings, broadcasts and eventually digital video. A demonstrator for the model was produced for sound recordings protection and complements that already developed for document delivery. The concept behind the model can be further developed so there should be even more opportunities to map this application onto different media and solve other problems.

3.10.1 Taking the Concepts of CITED Forward

From the CITED model other trials and research has grown until there are a whole host of models, hardware and software based on the concepts first explored by the CITED consortium (Cornish and Keates 1993).

3.10.1.1 The Virtual Workspace Concept

This issue was explored by the COPICAT project which addressed the area of electronic copyright protection by aiming to provide a basis for confidence in electronic copyright protection and open up a "blocked" market in multi-media electronic publishing.

COPICAT tried to develop a generic architectural model for an electronic copyright protection system incorporating the copyright-related event management model from the CITED project (ESPRIT 5469). It extended this by adding a security model appropriate to the application domain. Selected components from the EAST project (DELTA D2016) were to be used to create an educational copyright protection model. Selected multi-media educational material formed the basis for an example of material requiring copyright protection. The educational domain was chosen because the project consortium considered that this represented a "worst case" area in which most if not all copyright protection issues arise. Most other domains appeared to provide less stringent boundary conditions.

A complex ownership and access structure was simulated. The COPICAT system was installed and tested on a pilot site (University College Dublin).

A validation workpackage provided independent assurance of the effectiveness and correctness of specified aspects of the 6 key workpackages and delivered validation reports on a number of key deliverables. A novel feature of the validation was to subject the model to controlled and audited hostile attack by IT-experienced students.

The final security validation and verification report established that the technology developed in the project conferred an acceptable protection of rights and also that it was seen to do so by copyright owners.

COPICAT was important because it developed, but was unable to exploit, technology using the concept of the "virtual workspace" as described on the Project website at http://www.newcastle.research.ec.org/esp-syn/text/8195.html. In simple terms COPICAT developed a system whereby a work could be accessed and downloaded by any authorized person. They could then use the work for educational or research purposes so that it could be changed, or have information added or deleted. These actions are central to the "moral" rights referred to earlier. There was no question of economic benefit either to the owner or the user but only the facility to use the work as a basis for further development, study or experimentation. These actions were carried out in a virtual workspace which meant that. However, when the user tried to save the work this could not be done and the file simply disappeared. This model thus protected the integrity of the work and also prevented it being manipulated and changed and then subsequently re-issued as a different work by someone else. For a major electronic project to recognize the importance of moral rights and develop a system to protect them was a major breakthrough.

Other attempts to develop methods whereby a work could be accessed by tutor and student and subsequently changed failed largely because the technology used was too cumbersome and relied on telecommunications techniques rather than computer technology. An example of this is Project MURIEL (Website at http://www.cordis.lu/libraries/en/projects/muriel.html).

3.10.1.2 Trusted Third Parties (TTPs)

One of the great challenges for any electronic copyright management system is the collection of royalties. Fairly sophisticated technology now exists to control access, downloading, printing, changing text and obtaining data on how a product is used. The real issue for owners and users alike is how to pay for the use to which a work is put. There are a series of issues which need to be addressed in dealing with this complex and sensitive issue. Where a system is dealing with only one owner, or agent acting on behalf of a number of owners, then the issue is simply how to charge if charges are to be made. As discussed in another chapter, there are various models such as blanket licences, pay-per-view, pay-per-use, payment by

end-user, payment by parent institution, site licences, etc., etc. These can be arranged between the institution desiring access and the owner. The real problem comes when access to multiple works owned by multiple owners is necessary. It may be possible technically to allow access to a whole range of electronic materials through the same PC. This requires the use to switch from one supplier to another, using different ID and probably enjoying different access rights with each one. This is tiresome and irritating. Most researchers are unaware and uninterested in exactly which publisher or owner produces which piece of information provided it has the right academic or scientific status in terms of being refereed or guaranteed as to quality by the issuing institution. What most researchers want, and this is just as true of the advanced scientist as it is of the local historian working in the public library, is access to a wide range of materials with the minimum of protocols to observe. In the paper context users search for articles or reports with scant attention being paid to the publisher or originator and researchers will pass from one title to another without any such thought. This needs to be reflected in the electronic world as well. However, this raises major problems about how to pay the owner. What is needed is a system that will log use with all the necessary data, recover the royalties from the user (whether directly or via a billing system) and give the owner data on how the material has been used. Because there is no intermediary between user and owner and one use may need to relate to a number of owners the system breaks down unless some kind of "middle-man" is introduced rather like a warehouse in the commercial marketing chain. One role of such an intermediary is to collect data on the use of the material in question. This issue was tackled, amongst others by Project ERCOMS (Electronic Reserves Copyright Management Systems), a UK project which was led by De Montfort University in Milton Keynes. Details of the project can be found at http://www.iielr.dmu.ac.uk/Projects/ERCOMS/. The project identified that copyright monitoring and publisher feedback is a very important aspect of electronic library/reserve development. Under the various licensing schemes for networked texts, whether it be a fixed price licensing model, or usage-based or a combination of both, libraries and universities with networked systems have to demonstrate that they can control access and meter usage. The project managers realized that a generic electronic copyright management system which has built-in copyright management capabilities was thus urgently needed. De Montfort University aimed to develop such a system capable of working with various electronic reserve management systems, and able to provide full tracking of usage accountability and automatic counts of the occurrence of copyright events.

De Montfort University developed ERCOMS by building on the copyright and usage tracking experiences of other related library projects. The system was to be a complete package containing existing software and newly developed programs implemented on the server platform. Managers of electronic reserve systems would be able to feed in their rules on copyright management into the ERCOMS system, which would also provide an Application Programming Interface (API), a collection of programs to provide the linkage to the electronic materials for collecting the raw data. The data would serve to provide feedback to publishers on the use of their documents as well as providing evaluation data for project purpose. and The Open University Library.

The benefits of ERCOMS would include:

- A generic system which could be applied in other higher education institutions to any reserve system requiring copyright management functions;
- Reduced development time for setting up a copyright management system. ERCOMS would have all the major built-in functionality for copyright management. Reserve managers requiring copyright management facilities would not need to build a new system from scratch, but just feed their requirements into the system;
- A PC-based automated rights clearance system for handling electronic permission requests, generating chasers and user agreements.

Although using university reserve collections as the testbed for the system the application is capable of being mapped onto any similar situation where use data is required by a group of owners.

Unfortunately they once again designed a sensible model but was not implemented for economic and political reasons.

However, the struggle to find a solution to the Trusted Third Party problem persisted and was tackled in a different way, using a technology which is now regarded as commonplace but which, even five years ago, was still regarded with suspicion by many commercial factions – the Smartcard.

It was realized that the Smartcard, which can be used for so many applications today from personal data to paying for bus fares and gaining access to the local swimming pool, was likely to be a vehicle by which some of these problems could be solved. Building on the CITED experience, a team of researchers began to develop the idea of a smartcard to give access to electronic information and at the same time monitoring use and providing a mechanism for recording payments due. This led to project COPYSMART, The CopySmart project aimed to develop an industrial low-cost solution for implementing Intellectual Property Rights (IPR) management based on the CITED model.

COYSMART is interesting because it was originally driven by the media industry rather than the academic or research community. The Project recognized that the fast expansion of information networks like Internet and the introduction of digital broadcasting technology has given rise to the problem of media copyright, author rights, access control and payment for digital multimedia material. As already stated, once such material is published it is difficult to control the use, manipulation and distribution of digital information and to guarantee related rights. The project team realized that these issues need to be addressed for the development of the information society, for the creation of business and services on open networks and for the protection of the European cultural heritage. Thus it was driven by both commercial and non-commercial motives.

Although developed several years ago, the PC environment was then, as now, the one with the most urgent needs. The project began by looking at the fact that millions of PC users are connected to network servers and used, and to some extent still do use, removable storage media, such as CD-ROMs for leisure and business. Access to information on networks is mainly limited to free-ware or share-ware, because payment schemes are difficult to implement, copyright and author right protection even more. However, COPYSMART acknowledged that basic technology existed with:

- The CITED a global IPR management model previously described
- Technology from the Smart Card market, and
- Standard interfaces for portable hardware, the PCMCIA.

CopySmart targeted the PC environment and provided within short term the hardware and software building blocks for implementing IPR management in multimedia applications. Respect of standards was a key issue for large market acceptance. This issue of market acceptance is probably the biggest single barrier to developing a comprehensive ERMS that can be identified. The principle adopted was thus to not secure the PC, but the application and to provide hardware security by a standard PCMCIA (Personal Computer Memory Card International Association) device. This CopySmart device would contain protected CITED functions and security functions. Cryptographic algorithms and payment functions would reside in the CopySmart device in a removable security module, rather than in the PC itself, so that it can easily be adapted to national regulations and payment methods.

It was foreseen that an even larger and more demanding market that traditional publishing would be found in the TV environment with digital and interactive TV applications, but the issue of standardization had still to be addressed. The flexibility of the CITED architecture on which CopySmart relied and the respect of standard interfaces such as PCMCIA would allow re-use of CITED basic components for TV-centered applications in the future.

All the hardware and software developments were optimized for low cost implementation to facilitate dissemination and market acceptance. Wide dissemination was also to be encouraged by provision of development tools for integrators of IPR managed applications in the PC/Windows environment based on CopySmart building blocks. The project is described in more detail at the website http://www.newcastle.research.ec.org/esp-syn/text/20517.html.

CopySmart made strides in solving the basic problem of the TTP and distribution of royalties in return for use. Users obtain a CopySmart card which they then "charge" with the desired number of units. This can be done centrally or via the user's own institution such as a library. The user then uses the card to access electronic information and is debited as the use progresses according to the rates set by the rights owner. The usage is recorded and the revenue collected from charging the card is then distributed in proportion to the use made of different components of the electronic material accessed. The scheme is still in an embryonic state but the technology and concept are now both firmly established. What is needed is a critical mass of published material and owners so that installation and use of the system becomes worthwhile for the user and the institution. There is also a need for proper structures for collecting and distributing royalties in the same way as the traditional collecting societies.

One application for CopySmart so far has been in obtaining access to electronic materials for visually impaired persons. The software and hardware developed by Project SEDODEL creates, verifies, and demonstrates a secure document delivery service, which will meet the information needs of blind and partially sighted people, and guarantee the rights and obligations of actors in the publishing chain. It achieves this by integrating two key technologies: Electronic Rights Management Systems (ERMS) and accessible electronic documents. SEDODEL gives publishers the confidence to distribute electronic copies of their publications to organizations of and for the blind and partially sighted, and to blind and partially sighted people directly.

SEDODEL uses the CopySmart ECMS. CopySmart's end user software is implemented on a standard Windows PC with a smart card reader. Individual users have their own smart cards, which contain identification and authentication of the user together with the use rights granted by the service provider. A further EU-funded project under the Tide Programme, project SATURN has designed a set of smart card data structures for disabled and elderly people, which have been incorporated into a European standard.

The information is accessed by the end user on a CopySmart enabled PC running Windows. The end user's access rights are encoded on a personal smart card, which the application reads and interprets. The application unwraps the information and allows the user to access the information only in accordance with the specific user rights. The information will be accessed by a Document Reader, which will enable visually impaired users to read the information using their own access technology, such as screen readers, Braille displays and large print systems. Use of the information will be monitored by CopySmart, which deals with access control, clearing of rights, traceability, audit files, proofs of usage and handling of payments.

CopySmart achieved the implementation and operation of a secure document delivery service for blind and partially sighted people, the extension of CopySmart to the needs of blind and partially sighted people, and influenced changes in European copyright legislation. The secure service give publishers the confidence to release to organizations of and for the blind and partially sighted (and to blind and partially sighted people directly) electronic copies of their publications, thereby greatly enhancing access to information. Further information on CopySmart and its technical applications are described at http://www.snv.jussieu.fr/inova/ntevh/secure.htm.

3.11 No "All-in-One" Solution

It is clear that no one model solves all the issues facing the problems of making information available in electronic form.

Integrity was an issue tackled partly by COPICAT in that it provided the virtual workspace in which material could be used (or even abused!) without the original work being altered or the author's interests harmed because the subsequent text could not be saved for future use.

Payment for use was tackled by CopySmart with some success using smartcard technology. This has subsequently found a real development programme with Project SEDODEL. But neither CopySmart nor COPICAT could handle **both** issues and there would need to be a conflation of the two projects and their hardware/software to achieve a solution which covered more than one issue.

3.11.1 Copying, Re-Use and Re-Transmission

None of these projects in themselves even attempted to solve the crucial issue of preventing downloading, re-transmission and subsequent use by unauthorized users. Many systems have been tried but none have so far proved totally effective. Whilst copying is an irritant, provided that it takes pace either on to paper or even into a stand-alone PC it is rarely more than that except in the "pop" music industry where it seriously undermines sales which are primarily direct to individuals. This is why the recent controversy over MP3 and Napster (Napster 2001) has been so furiously fought. However, in the scientific information world, there are two problems: downloading and retransmission, These can be considered either in terms of prevention or identification of unauthorized copies. In the first instance a technical mechanism has to be put in place that will prevent anyone downloading or re-transmitting an electronic document unless they have permission to do so. Such permission could be in terms of a licence or other permission from the owner direct or through some kind of permission given to an intermediary such as a library. The second option, identifying "copies" made and re-transmitted has the advantage that it is technically possible but can only be enforced after the event and therefore will often fail to find infringing copies or unlawful use except by accident or very complex tracking mechanisms. What is essential is that there is a consensus between owners, creators and users as to what is needed, what is desirable and what is possible in both technical and legal terms. Such a consensus was the aim of Project IMPRIMATUR.

3.11.2 Exceptions for Users

Although this may seem a minor issue, the ability of individuals to access information without being inhibited either by technology or economics is an essential element in the information flow which is vital to human development, education, economic growth, scientific research and democracy and justice. The crucial question is how to manage the se exceptions electronically. In a paper world there is nothing to stop occasional copying by individuals anyway but once material is digital then the opposite situation applies and the copyright owner can easily put in place mechanisms which can prevent totally any access unless the pre-conditions set down by that owner are met in full. The technology therefore is in danger of working against the needs of the individual and benefiting only the owners of copyright material. The unsolved challenge is how to design a protection mechanism that will nevertheless allow certain amounts of use which are considered reasonable within the law without either reference to the owners or the need for payment. This was perhaps the greatest single matter for debate when the EU Directive on copyright was passing through its various stages. At one point the exceptions in favour of individual users were being ignored because it was felt that the electronic environment provided for total management. But pressure groups such as libraries (for example the European Bureau of Library, Information and Documentation Associations (EBLIDA)), as demonstrated on their website www.eblida.org eventually succeeded in inserting protection for users through Article 6(4) which states:

> Notwithstanding the legal protection provided for in Paragraph 1,in the absence of voluntary measures taken by rightsholders, ….Member States shall take appropriate measures to ensure

> the rightsholders make available to the beneficiary of an exception or limitation provided for in national law the means of benefiting from the exception or limitation…where the beneficiary has legal access to the protected work or subject-matter concerned.

What has not yet been resolved is just how such exceptions can be managed. Is it possible to tell an electronic management system that one act of downloading is allowed by law but exactly the same action, undertaken for a different purpose, is not and should be controlled by the owner or paid for by the user. There are many issues which will take a long time to resolve. Rightsowners tend to say that they are reasonable and will allow many of the actions permitted by law anyway but even then it is hard to see how they can be distinguished one from the other. Users fear that, whilst rightsowners may intend to be reasonable and cooperative now they may not be so in the future. The installation of mechanisms that control, even if no fee is sought, are seen as the beginning of control of legal exceptions by owners and therefore this model is being fiercely resisted by user groups.

4.0 DEVELOPING A CONSENSUS

As has already been demonstrated, the whole area of electronic rights management is one of considerable potential conflict between creators, distributors (whether publishers, database hosts, website providers) library and information professionals and end-users.

Being aware of these potential conflicts the European Commission funded a major project – IMPRIMATUR (Intellectual Multimedia Property Rights Model And Terminology For Universal reference) – to try to build this consensus amongst the major players in the EU, Japan, Australia and North America. Essentially IMPRIMATUR was considered a "horizontal" project, in other words it worked across a number of disciplines and areas rather than within any particular one. The fact that the subject of copyright management crosses more than one topic and discipline boundaries has already been demonstrated in this paper. As well as the more obvious players it also is crucial to bring together writers, photographers, composers, artists, film makers, record makers, recording artists, performers, lawyers, software and hardware manufacturers, Internet providers, librarians, users and financial specialists. Copyright knows no subject barriers any more than the human mind does. Because copyright is essentially a commercial product with a market value. Project IMPRIMATUR worked in the context of electronic commerce and not just another forum for agreeing how to manage copyright. For this reason the consortium which originally ran IMPRIMATUR consisted of quite a large number of partners representing many of these aspects of the information industry. It included a national society representing authors rights and the international organization representing authors and composers collecting societies, technical expertise, electronic banking specialists, academics, users and intermediaries, the entertainment industry and legal experts. The idea was to build a consensus on what should be managed and how without stipulating the technical mechanisms which would be seen as antitrust (in the US) and anti competitive.

The consensus processes provided a framework for technological development as it provided a commercial and legal context for this development. The aim of the technical development was to integrate, demonstrate and validate the key features of an Electronic Right Management System.

By developing a prototype technology the Project tried to move forward the debate by offering practical solutions to real copyright management problems. The hardware and software developed by IMPRIMATUR was tried out major photographic archive in the University of Florence where an image can be sought and found, identified for potential use, requested online with the purposes defined and the image received with the necessary watermarking and payment records. This can be achieved in a matter of two or three minutes. The technology has also been used successfully protecting sound files. The technology can be used to adapt and develop the technology to meet specific media providers of all kinds as well as end-user requirements. The watermarking technology developed by IMPRIMATUR is

transparent to the user but is also permanent and therefore enables owners to identify and trace infringing copies.

All this is very exciting and challenging but needs to be set in the context of the whole range of interested parties. The project took great trouble to hold meetings of Special Interest Groups (SIGs) in areas of interest (sound recording industry, libraries, electronic commerce) These were supplemented by major consensus for a general publicity and the establishment of official contacts and networks.

One example of working together rather than in a vacuum is the question of identifiers. Although electronic technology can do many things, it requires the facility to identify the things to be managed. One initiative, the Digital Object Identifier (DOI) comes from America. This is a unique and persistent identifier to mark digital objects in a global electronic environment. It is managed by a directory to link users to a specific content whose ownership is recorded centrally. However, ownership is only one half of the problem. Having gained access to a work and knowing who owns it, what can the user do with it and under what circumstances. There is also the question of a multiplicity of rights in one work, especially one which contains sound, moving images and text as well as photographs and graphic works. To cope with these complex issues the International Confederation of Authors and Composers Societies (CISAC) developed an identification system which includes a wide range of works which can be tagged to give all creators and parties with an interest in the work as well as licensing conditions. Once matched with DOI this could provide a world-wide access and rights management database.

No project has all the solutions because nobody is quite sure of all the questions. Technically it is possible to build an “all-singing, all-dancing” electronic rights management system but the question has to be asked whether it is really needed. There are serious issues as to whether such a system could ever be economical or realistic. There are serious questions from a scientific point of view as to whether it might threaten the information flow on which everyone is all dependent for professional existence. In the academic and text area it may well prove too expensive to build and maintain such systems. In the multi-billion dollar entertainment industry the argument is quite different. Once a feature film, such as *Harry Potter* becomes available in digital form the potential for infringement in a massive scale gives a real reason for building such systems At the same time it gives the owner immense potential for exploiting a work in this form. Different elements of the film can be digitally tagged and use and access to them managed accordingly.

But once the technology has been developed, as in all other cases, the cost will diminish and implementation may become quite easy. On the other hand, managing the rights which the technology gives may continue to be too costly and burdensome to make it worthwhile. The economics of protection has not yet become sufficiently stable to be able to determine the future. As one protection device succeeds another, other issues now arise. For example, CD manufacturers have devised blocking technology to prevent their products being played on PCs. Or, indeed, making them unplayable in this situation. Purchasers of the CDs have taken this issue to court as it infringes their rights under consumer protection legislation in some countries.

For example it was reported that Universal, the world’s biggest record company, was to release more copy-proof CDs after music sales declined for the second year running. They were down by 5% in 2001. One problem is the availability of CD rewriters which have enabled music listeners to copy CDs cheaply and easily. IFPI (International Federation of the Phonographic Industry) members closed down 1,000 illicit music internet sites last year but there are still many remaining (IFPI 2002).

The development of consensus building is continuing through the company set up as a result of IMPRIMATUR called Rightscom. In their own words from the website www.rightscom.com the organization itself says:

> In the digital community, no company or organization can work in isolation from the wider environment. Conformance to international standards is just one area where building

industry-wide consensus has become a prerequisite. Just as important is obtaining "buy-in" within large organizations. For sustained development in the digital world, consensus building is now an essential management task. Two aspects of the digital market make consensus building both complex and challenging. The interdependency of the issues which cross traditional organizational borders can rapidly lead to confusion. Secondly, stakeholders themselves possess varying levels of knowledge, and often have widely divergent and conflicting expectations. Rightscom has both the expertise and the sensitivity needed to seek out solutions on which all stakeholders can agree, promoting positive and constructive outcomes for projects. Rightscom's work in consensus building initiatives has led to the development of a unique and extremely effective methodology.

5.0 CONCLUSION

Clearly there is much still to be done. No system has found the complete answer to the problems of promoting, yet protecting information in electronic form. Economics, politics, legal issues and consumer resistance may in the end determine how these issues are resolved. Access, integrity, paternity, printing, downloading and royalty payments can all be managed. Distribution of royalties is causing difficulties still and nobody has yet produced a solution for managing the exceptions to copyright in favour of users. The future is certainly challenging as both law and technology develop. Similarly, as users become more aware of the possibilities of information delivery their expectations will change and fundamentally alter attitudes to intellectual property in a world where every user may well become an owner.

6.0 REFERENCES

Cornish, G.P. (1992). Copyright protection for materials in electronic format: the CITED solution. *Learned Publishing,* 6(2): 15-18.

Cornish, G.P. (2000). Looking both ways: the challenge to the intermediary in an electronic age. In: Connoly, P. & Reiddy, D. (eds) *The digital library: challenges and solutions for the new millennium. Proceedings of an international conference held in Bologna, July 1998,* pp. 29-38.

Cornish, G.P. & Keates, S.L. (1993). Copyright protection of artistic works in electronic form: the CITED approach. *Information services and use,* 13(4): 389-398.

European Parliament. (2001). Directive 2001/29/C on the harmonization of certain aspects of copyright and related rights in the information society.

Feather, J. (1994). *Publishing, piracy and politics,* London: Mansell.

IFPI. (2002). International Federation of the Phonographic Industry. *Guardian* (newspaper) 17/04/02 p. 20.

Napster. (2001). Copyright charges upheld. *USA Toda,* 26 June, 2001.

Renwick, K. (1991). Booksellers and agents – help or hindrance. *IATUL Quarterly,* 5(4): 245-255.

Van Slype, G and Van Halm, J. (1988). *Evaluation of experiments in electronic document delivery and electronic publishing.* (EEC, EUR report 11208).

Infrastructure of Electronic Information Management

Gregory D. Twitchell *and* Michael T. Frame
U.S. Geological Survey
12201 Sunrise Valley Drive / MS 302
Reston, Virginia 22092
USA

Gregory_Twitchell@usgs.gov, mike_frame@usgs.gov

ABSTRACT

The information technology infrastructure of an organization, whether it is a private, non-profit, federal, or academic institution, is key to delivering timely and high-quality products and services to its customers and stakeholders. With the evolution of the Internet and the World Wide Web, resources that were once "centralized" in nature are now distributed across the organization in various locations and often remote regions of the country. This presents tremendous challenges to the information technology managers, users, and CEOs of large world-wide corporations who wish to exchange information or get access to resources in today's global marketplace. Several tools and technologies have been developed over recent years that play critical roles in ensuring that the proper information infrastructure exists within the organization to facilitate this global information marketplace Such tools and technologies as JAVA, Proxy Servers, Virtual Private Networks (VPN), multi-platform database management solutions, high-speed telecommunication technologies (ATM, ISDN, etc.), mass storage devices, and firewall technologies most often determine the organization's success through effective and efficient information infrastructure practices. This session will address several of these technologies and provide options related to those that may exist and can be readily applied within Eastern Europe.

1.0 NETWORK INFRASTRUCTURE

There is a need in today's environment to provide high performance, scalable, and robust systems to almost all businesses, regardless of size. A strong network infrastructure is the key to providing reliable systems with minimal downtime for mission-critical applications. A major component of insuring data integrity is the provision of network security. Gone are the days of restricting physical access and implementing password protection to insure data integrity. That is just a small component in the overall scheme. With the explosion of the Internet, access attacks on networks can occur from anywhere and at anytime. With that type of exposure it is critical that networks are protected not only from hardware and software failures, but also from cyber attacks. The following will provide a general overview of network infrastructure.

2.0 MASS STORAGE DEVICES SYSTEM FAULT TOLERANCE

2.1 Random Array of Independent Drives disk system (RAID)

The Random Array of Independent Drives disk (RAID) system has the capability to protect data and remain on-line with data access despite a single disk failure (RAID storage systems with two concurrent disk failures can continue to operate with a hot disk standby). Once the failed drive(s) is replaced, the system will rebuild the hard drive while remaining online. RAID system can be supported in both hardware or software configuration.

2.2 Disk Mirroring

This is the most elementary type of disk system that provides for system fault tolerance. It requires two hard disk drives. These two drives are duplicates of each other. If there is a failure of one of the drives, the disk system will continue to operate and provide on-line data access. When the failed drive is replaced the system will provide for an on-line reconstruction of the new drive by copying the entire contents of the operational drive to the newly installed replacement drive.

2.3 Extended Data Availability and Protection (EDAP)

Extended Data Availability and Protection (EDAP) is a storage system that provides data protection and access to data, despite failures within the disk system, or, any attached systems, or environmental failures. RAID is considered the lowest level of EDAP.

EDAP techniques for disk are designated as RAID Levels 1-5. Mirroring and Parity RAID are two types that provide various levels of EDAP for disks. The disadvantage of mirroring is that it requires 100% redundancy, while Parity RAID read performance is enhanced. Impact on write performance for Parity RAID is modest. Additionally when failure occurs a higher percentage of a mirrored group may fail in relation to Parity RAID.

RAID levels 3, 4, and 5 are usually identified as Parity RAID. In the RAID Level 3 array, all the disks operate in parallel. RAID Level 3 is useful for high bandwidth applications while RAID Levels 4 and 5 are more suitable for high transaction rate applications. In comparison to the 100% redundancy required by mirroring, RAID levels 3-5 requires only 10% to 33%. The differences in the RAID levels 3-5 are related to how the data and the redundant data are mapped to the disk. EDAP operates in three states:

- Normal (Protected) – EDAP capability is not being employed to counteract a failure.
- Reduced (Vulnerable) – EDAP is in use to counteract a failure of one disk.
- Down – A state in which data cannot be stored or retrieved.

In a normal state, EDAP is running at peak performance, providing on-line storage and retrieval of data. Performance is affected during the reduced state due to regeneration of data from the failed disk and rebuilding of the failed disk. This regeneration of data requires additional processing time and impacts the overall I/O performance. EDAP design is critical to minimize impact for the performance level during the storage system's reduced state.

In addition to maintaining on-line access to reliable data during a failure, EDAP can minimize the time period that the storage system is in a reduced state by supporting on-line sparing of disks. This will facilitate the reconstruction of the reliable data immediately upon a disk failure. EDAP can also minimize the period in which a storage system is in a down state by providing hot swapping of failed disks. This would make it possible to replace disks without powering down the system, therefore putting them in a down state.

2.4 RAID System Fault Tolerance Systems

Protection of computer data from anomalies, such as human errors, hardware failures, or environmental conditions has been a top priority since the development of the first computer system. A strong data backup strategy is a primary and essential component to insure data protection. Advances in hardware reliability have minimized loss of data due to hardware failures and environmental conditions, but minimizing human failures has been more difficult.

Historically, “mirroring” the data was first level of protection, in addition to maintaining the immediate access to data in the event of a failure of one of the disks in a mirrored pair. Implementation of a mirrored system was twice as expensive as a non-mirrored system. The need to develop a less expensive system to protect data and on-line access led to the development of RAID (Redundant Array of Independent Disk Drives).

Early Parity RAID disk systems suffered from poor performance due primarily to the need to generate and write parity during a write operation and to regenerate the data on-line in response to I/O requests for data from a failed disk. This performance problem was addressed by the use of caching and write-assist disks. Different RAID levels basically describe how data and redundant data are mapped across the disks of an array.

Parity RAID requires a minimum of three physical disks. One disk is dedicated to parity, a second to the first data set, and a third disk to the second data set. Data and parity can be configured to map to disks in a manner so that one disk is not totally dedicated to parity.

2.5 Redundancy Performance

Mirroring or Parity RAID that provides redundancy against a disk failure also provides performance advantages. Multiple disks in an array can be used in parallel for applications requiring high transfer rates (bandwidth) or independently for applications requiring high transaction rates (asynchronous I/O requests involving relatively small amounts of data for each I/O task).

Controller, device channeling and redundancy in disk may enhance disk system performance. Performance can be improved by having all components in a redundancy group actively on-line to share I/O tasks. If any one component fails, the performance will degrade until the defective component is replaced.

Table 1: RAID Level Summary

RAID Level	Description	Data Reliability (protection)	Data Transfer Rates	I/O Rates
0	Striping of data across multiple drives in an array. This is a high performance solution, however there is no data protection.	No data protection	Very High	Very high for both eads and writes
1	Raid 1 is known as mirroring. Mirroring is the 100% duplication of data from one disk to another. This is a high availability solution, but due to the 100% duplication, it is a costly solution.	Excellent reliability	Reads are higher than a single drive. Writes are about the same as a single drive.	Reads are up to two times faster than a single drive. Writes are about the same as a single drive.
5	This is the most widely deployed RAID level. This level provides a balance between performance and cost. Striping with parity. Data and parity information is spread among each drive in the drive group. Parity is equal to the total number of disks in the volume minus one drive.	Good reliability	Reads are similar to RAID 0. Writes are slower than a single drive due to penalty of writing parity.	Reads are similar to RAID 0. Writes are usually slower than a single drive.

3.0 JAVA

Security features provided by Java™ are intended for a variety of audiences, including end users and developers.

For users there is built-in security that prevents malicious programs such as viruses from running, while maintaining privacy about their files and any information about them. In Java™ 1.2 security controls can be invoked when desired for applications, similar to those for applets in previous versions.

Developers can use application-programming interfaces (API) to invoke security for programs. The framework for API enables administrators to define, and then integrate, security to control access to resources. These include authentication and authorization services, cryptography service, security manager service, and policy implementations. Java™ Authentication and authorization services (JAAS) provide support that administrators can define by users, groups or roles. Java™ Cryptology Extension (JCE) allow for encryption, key generation and agreement, and message authentication and code. JCE also allows for the addition of other qualified cryptography libraries. To allow secure Internet connections Java™ Secure Socket Extension (JSSE) packages include Java™ versions Secure Sockets Layer (SSL) and Transport Layer Security (TLS) protocols. Use of JSSE enables HTTP, Telnet, NNTP, and FTP and ensures the security of the data passing between the client and the server.

Users can also manage public/private key and public key certificates from people they trust. Java™ tools allow management of database of keys and certificates, digital signatures for Java ARchive (JAR) files, authentication of signatures and integrity of content. Users can create and modify policies files that will allow them to define and control their environment.

These policies have evolved from the original Java™ security model, known as the “sandbox” model. In this model local code is given full access to system resources (i.e. a file system). Applets that are downloaded would have only limited access to system resources with the “sandbox.” Permissions would include checking for file existence, read rights, write rights, renaming rights, directory rights creation, listing of files, file type, file size and file timestamp.

4.0 PROXY SERVERS

Browsers such as Internet Explorer and Netscape allow the user to browse the web without restrictions. The client is free to request and access any web page without regard to its content. Due to this flexibility the user may be able to access web information that is inappropriate for certain situations.

Proxy servers allow administrators to limit access to certain content within a network environment. A proxy server resides between the client or end user and an external server on the World Wide Web. The placement of this server within that environment determines its control over an entire domain or a group of individuals. The proxy server resides strategically on the firewall and will intercept all of an end user’s requests at the firewall. If a requested web page is not restricted by access control list (ACL) the proxy server processes the request and the web page is sent to the client. However, if the requested page is on the ACL, the client will receive a message that the web page is not valid or not accessible.

A typical network configuration places the proxy server as the Internet gateway for users whose access to the web is restricted.

Internet performance can be enhanced with the use of a proxy server if it functions as a caching server. The proxy server can cache web pages previously requested by users without the need to go the Internet. For example if a number of users request the same web site, the first user's request will go to the web, the page will be downloaded to the user and the proxy server will write the page – cache that web page data – to its hard drive. Any subsequent request for that web page is served from the proxy server's cache. This will avoid unnecessary duplicate requests and delays that might occur from the Internet. However, with the explosive growth of the web, maintaining the ACL for proxy servers is an administrative function that can be very labor-intensive.

Proxy servers cannot be a total solution for controlling inappropriate or objectionable material from getting to the end user. They will not prevent inappropriate material in an email attachment, nor will they filter transmission of objectionable material in a chat session. Most proxy servers can accept domain names; however, controlling inappropriate pages from downloading is a difficult task.

Proxy servers strengths are their provision of a higher level of control than exists with end users using unrestricted browsers, and their ability to process access to web pages more efficiently.

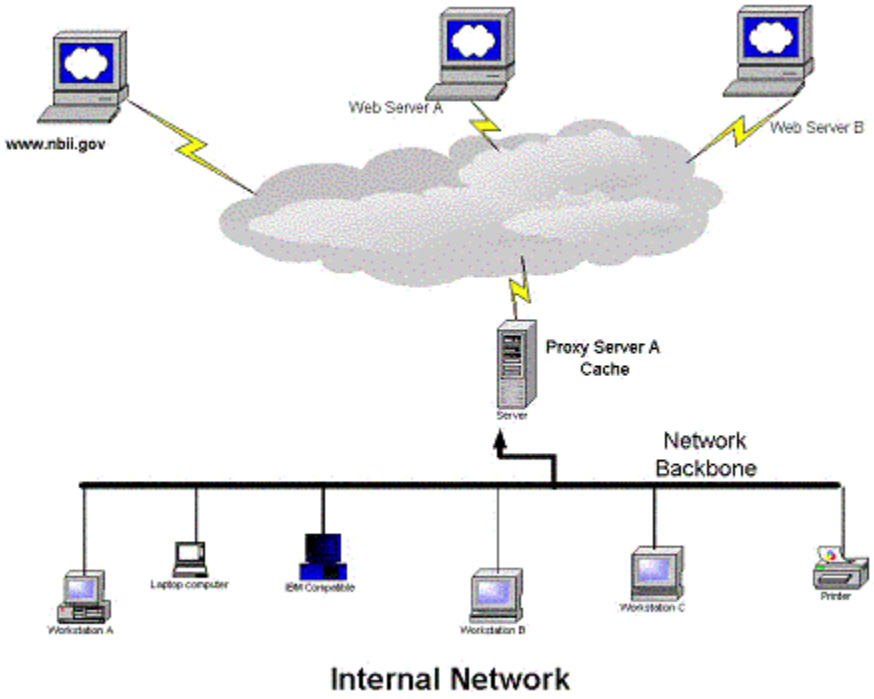


Figure 1a: Proxy Server.

Workstation A initiates a request for www.nbii.gov. The request is routed through Proxy Server A, which processes the request and returns the web page to Workstation A (i.e. assuming it is on the approved ACL). The web page is also written to Proxy Server's A cache (i.e. hard drive). Any subsequent request for www.nbii.gov from a workstation on the internal network, the proxy server will deliver the information from its local cache. The end result is network performance is enhanced by not requiring a search for www.nbii.gov outside the internal LAN, which reduces overall bandwidth usage.

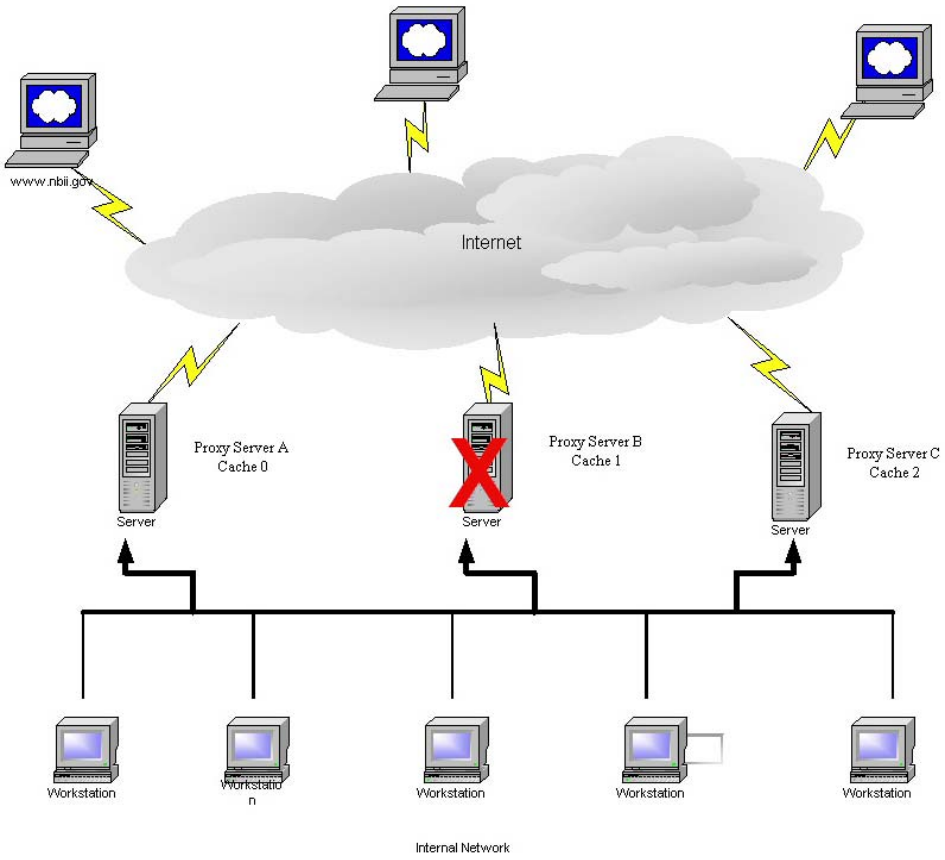


Figure 1b: Multiple Proxy Server Environment.

The internal network workstations will request information from the Internet. On the initial request the proxy server will retrieve that information and serve it to the internal workstation and cache it locally. For a subsequent request for the same information by another internal workstation the proxy server will serve that information from its local cache rather than retrieve it additional times from the Internet. The loss of Proxy server B does not effect network performance, because of server redundancy. The cache on all three servers are identical and any request handled by server B is simply rerouted to server A or C. Proxy servers can also use Network Address Translation (NAT).

5.0 NETWORK ADDRESS TRANSLATION (NAT)

The limits on the availability of IP network addresses in IP version 4.0, coupled with the explosive growth of the Internet, have resulted in an inadequate number of unique IP addresses to meet demand. One solution to this problem is the use of Network Address Translation (NAT).

NAT is able to translate IP addresses by setting up a transition table of all internal IP addresses that will send data packets through the NAT router. With the use of NAT the external world will see only a limited number of valid registered IP addresses. Internally to the organization a private addressing scheme could support a large number of network hosts. Each interface that leads to an outside interface would have a valid registered

address. This is an advantage from a security perspective. Any internal host will be unable to receive an incoming IP connection from an external system unless the external interface (i.e. gateway) is specifically configured to allow the connection. In order to establish integrity in the internal network all external interfaces must be configured with NAT.

NAT exists in two modes; static NAT and dynamic NAT. Static NAT maps internal IP addresses to external IP addresses on a one-to-one basis. Dynamic NAT maps all internal IP addresses to use one external IP address.

The recommended operation to NAT is detailed in Request for Comments (RFC) RFC 1918, which describes the recommended private internal addressing schemes. Standards for NAT are established in RFC 1631.

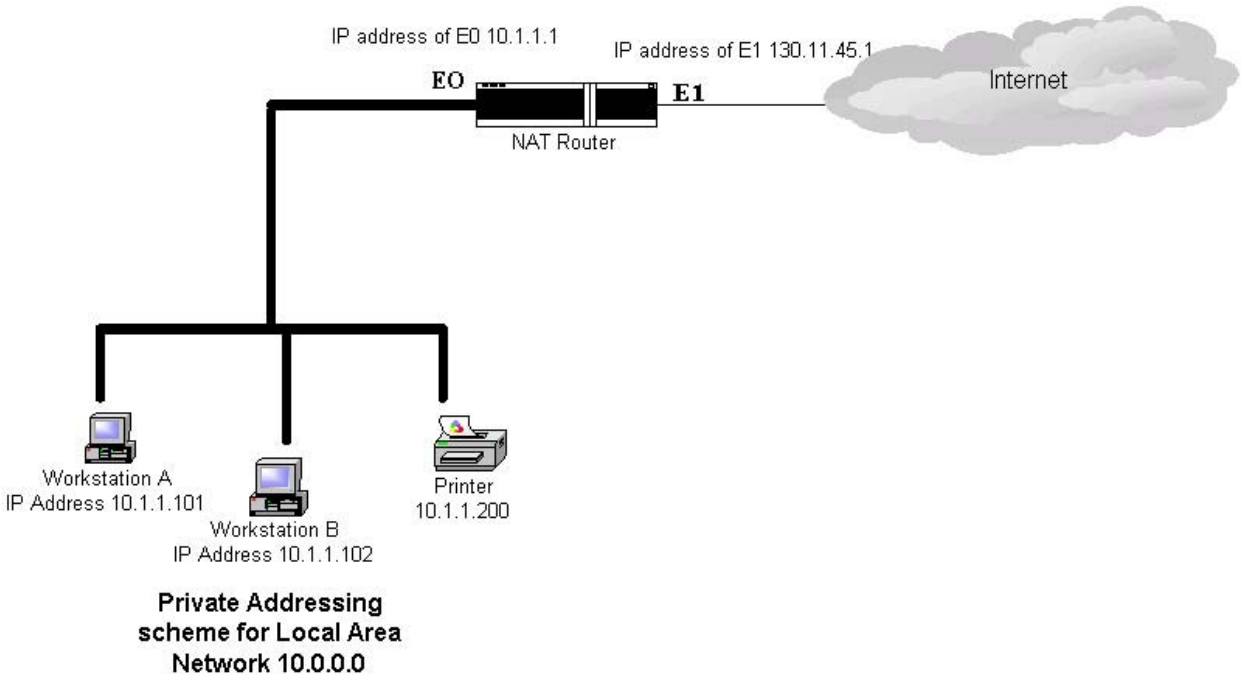


Figure 2: Network Address Translation (NAT).

This figure shows a NAT enabled router with a network address of 10.1.1.1. When any host on the 10.0.0.0 internal LAN makes a request to an external host (i.e. the Internet) NAT will translate the 10.0.0.0 addresses to 130.11.45.1. The internal hosts can access any host on the external network. For the external hosts looking in it has the appearance that all inbound and outbound data are originating from the single IP address of 130.11.45.1 (i.e. E1 the route). Figure 2.3 is an example of NAT in the dynamic mode.

6.0 FIREWALL

A firewall is a security host sitting in between a company's internal network and the external world (Internet). Firewalls are usually the company's first line of defense to prevent attacks from the outside. There are different types of firewalls in use today. The two most popular are Packet Filtering and Proxy Servers. These types of firewalls have different capabilities. Firewalls that filter data packets will allow or deny the entry of

traffic based on the IP address or source and destination ports. Proxy firewalls are based on specific applications used, including http, telnet, ftp and ssl traffic. The firewall will check this type of traffic according to the specific rules that are defined for those applications. The organization's security profile will determine the type of firewall used.

An organization that has an Internet connection should install a firewall for two primary reasons. First, company data must be protected by a malicious attack. An attack could be in the form of malicious code (i.e. virus attack) or an individual hacking into the network. Downtime from this attack could result in loss of productivity, increased expenses and revenue loss. While backups can restore the data, there is potential for stolen data. The value of the data that could be compromised is the second reason to install a firewall. The data that is stolen could contain confidential company data. In many cases, stolen data maybe used against a company. For these reasons, firewall implementation is just the first step to protect data.

6.1 Intrusion Detection

Without the proper monitoring tools in place an intruder may bypass a company's firewall. The damage that can be done could be severe if the organization is not prepared. Some skilled hackers can get through certain firewalls without notice. Deploying intrusion detection tools at the perimeter of a network is just as important as deploying a firewall. These two technologies must be used in tandem at network boundaries; a company would want to know if a hacker penetrates the firewall and gain access to the network. The use of these tools enables real-time monitoring and alerts to the appropriate individuals. With these alerts, necessary actions can be taken to protect valuable data. As an extra precaution companies should deploy a firewall and intrusion detection tools between most departments in the organization.

6.2 The Insider Attack

Hackers are not the only threats to a company's valuable data; many security problems can be directly attributed to their employees. Training the end user is a very important action for a company to take; after all, the end users have access to some or all of the company's data. Investment in training the users is necessary so accidents do not occur and company data is not compromised. If internal attacks do occur, intrusion detection software can monitor and track the electronic activities of the employees.

6.3 Firewall Conclusion

A combination of tools is necessary in order to protect data internally and externally; one tool is usually not enough. Combining various technologies such as firewall and intrusion detection software will maximize the ability to monitor and protect valuable data.

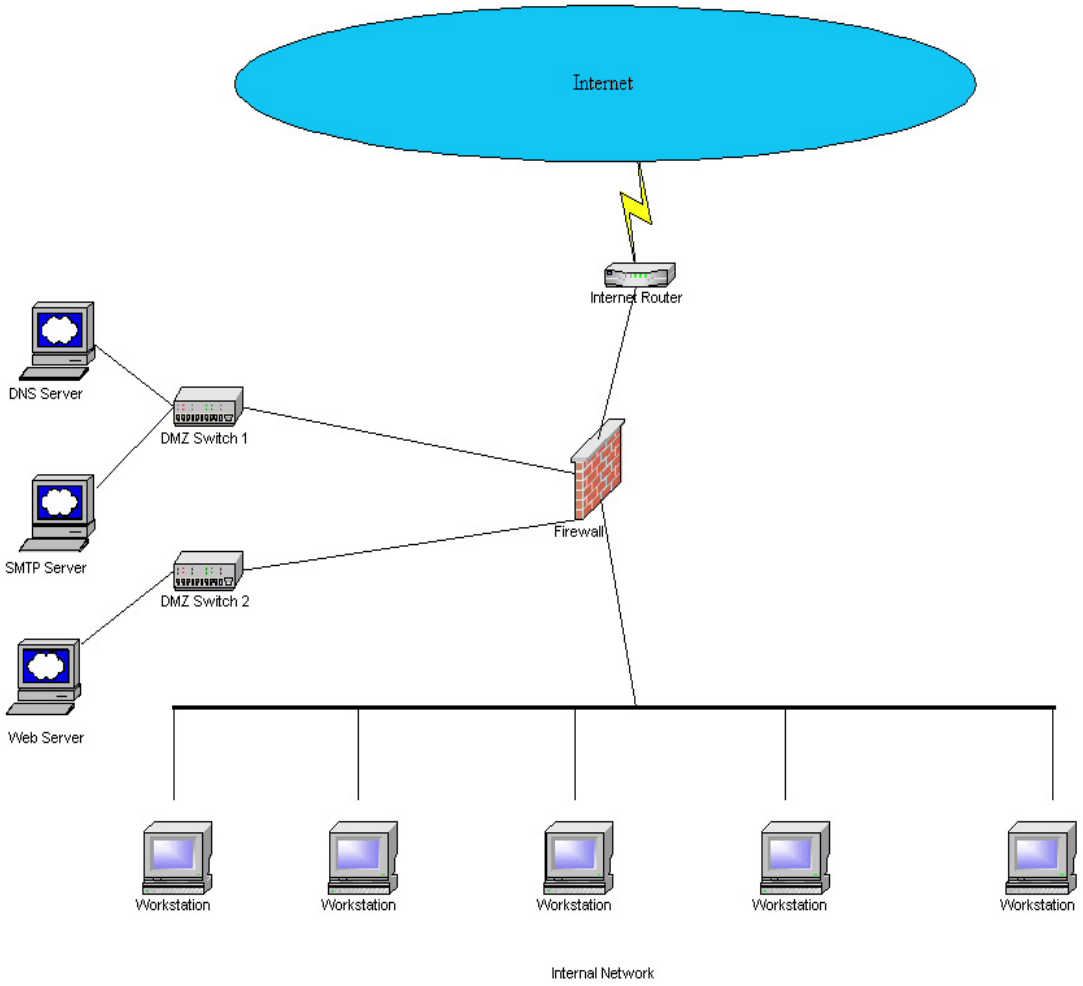


Figure 3: Firewall.

An example of a LAN with a firewall that provides filtering of content for inbound and outbound data. A DMZ on a protected leg allow external users to have read access, but prevents initiation of any requests to and fro internal LAN access.

7.0 VIRTUAL PRIVATE NETWORKS

Virtual Private Networks (VPNs) are becoming a popular way to deploy private networks across a wide geographic area. A VPN device can involve either hardware or software; installation on the server (sender of information) or the client (recipient of information) may be necessary. These devices are used to establish a secured session between the server and the client. Virtual private networks use a public network to link one or more endpoints. An endpoint can be a network device, such as a router or a user, such as a personal computer. If the endpoints are two network devices, that communication is considered a LAN-to-LAN VPN connection. For example, an organization's Virginia office network could connect to its Denver office network over a LAN-to-LAN VPN connection. If the endpoints are an end user's personal computer and a network device,

the communication is a LAN-to-network client VPN connection. A traveling user needing to connect to his LAN to read his email is an example of a LAN-to-network client VPN connection. This communication link exists only when a session is active; once completed, the session is terminated and the link is destroyed.

The usual mode of connection is for the traveler is to connect in a dial up mode to an Internet Service Provider (ISP); the software would then create a secure VPN session with the mail server (See Figure 4 for a graphical representation of a virtual private network).

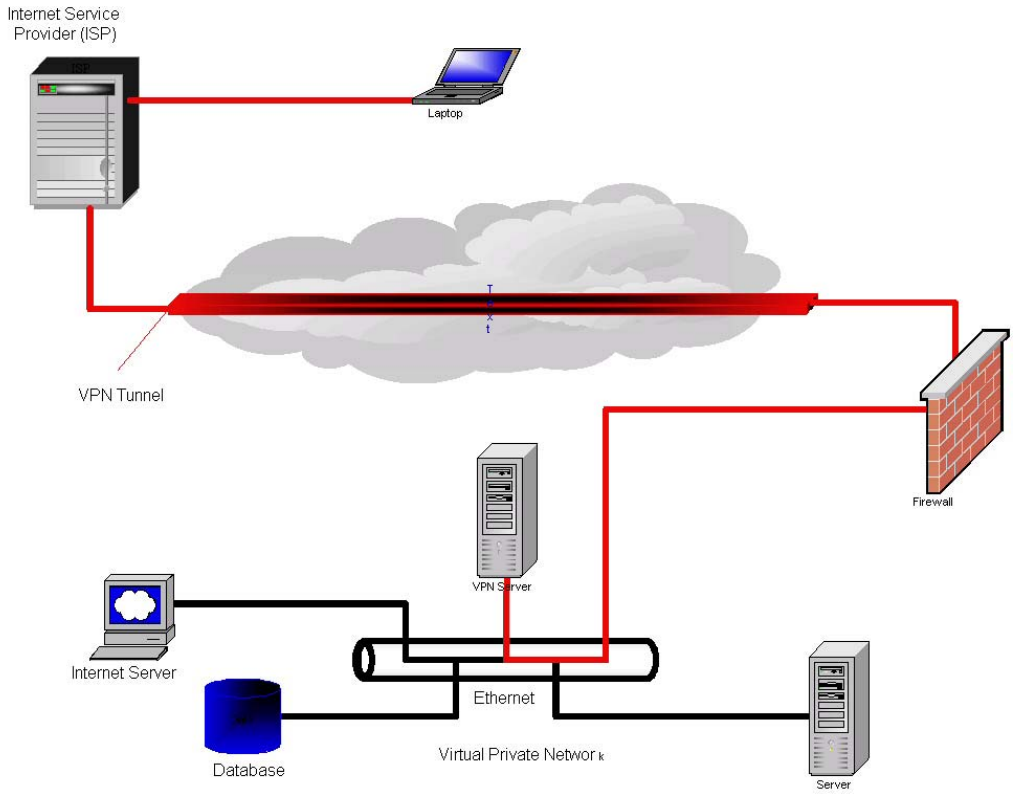


Figure 4: Virtual Private Network.

With rapidly changing technology there are a variety of ways and options for implementing VPNs.

In this example, a remote laptop user dials into an Internet Service Provider (ISP). Once the connection is established, the remote user runs the local VPN software. The software will build a secure VPN tunnel through the Internet and the user is authenticated by the LAN's VPN server. Once authenticated by the VPN server, the remote user has access to the LAN resources. Once the user session ends the virtual path is destroyed.

7.1 Point-to-Point Tunneling Protocol (PPTP)

One of the most widely used VPN protocols in use is Point-to-Point Tunneling Protocol (PPTP). PPTP encapsulates a Point-to-Point (PPP) frame. PPP was first widely used to dial into a remote network

using a modem. PPP transmits a network specific packet by encapsulating it into an IP (Internet Protocol) packet for transmission over the Internet. This enables the transfer of non-routable protocols such as IPX (Internetwork Packet Exchange), NetBeui, and AppleTalk, in addition to TCP/IP (Transmission Control Protocol/Internet Protocol).

PPTP has many of the characteristics of PPP, because it is very flexible in its ability to be used for non-TCP/IP environments. PPTP is widely used due to the popularity of the Windows operating system. The Microsoft Corporation developed PPTP; it is a widely deployed protocol used in Windows 9x/ME, Windows NT, Windows 2000, and Windows XP. This deployment of client software allows for the use of voluntary VPNs. PPTP authentication uses Password Authentication Protocol (PAP) and Challenge-Handshake Authentication Protocol (CHAP). Microsoft has developed its own PPTP authentication called MS-CHAP; this uses NT Domain information for authentication, which is based on RC4, based 40-bit or 128-bit encryption. MPPE is Microsoft's implementation of its PPTP client software and is the solution for voluntary mode access.

7.2 Layer 2 Forwarding Protocol (L2F)

Layer 2 Forwarding (L2F) has many similarities to PPTP. Like PPTP, L2F was designed to work with PPP and support non-routable protocols. Additional benefits of L2F over PPTP are its ability to support more authentication standards, differing types of networks and multiple threads on a single connection. The additional authentication standards include Terminal Access Controller Access Control System (TACACS) and Remote Authentication Dial-in User Service (RADIUS). Both of these standards authenticate at the beginning of the session transmission. Frame Relay and Asynchronous Transfer Mode (ATM) are examples of different types of networks that L2F supports. PPTP allows only one client to connect over a single connection, whereas L2F allows for multiple connections over a single tunnel.

7.3 Layer 2 Tunneling Protocol (L2TP)

Layer 2 Tunneling Protocol (L2TP) is used primarily in mandatory mode to access VPNs. It has the same capabilities as L2F and PPTP. Like L2F, L2TP supports multiple connections through one tunnel, works with various types of networks, and supports non-routable protocols (i.e. both IP and non-IP traffic). Unlike the others however, L2TP is IPSec compliant (see Section 2.4), which provides for stronger encryption standards, authentication and key management. The sender of any packet is able to encrypt and/or authenticate each packet. Encryption and authentication of packets leads to the use of two modes, transport and tunnel mode. In the transport mode only the transport layer is encrypted or authenticated. In tunnel mode, encryption and authentication are applied to the entire packet, not just one segment. The tunnel mode method provides for the greatest protection against attacks due to the increase in security.

7.4 IPSec Encryption

IP Security protocol (IPSec) encryption is an Internet Engineering Task Force standard that supports 56-bit and 168-bit encryption algorithms in client software. IPSec uses two protocols: AH (Authentication header) and ESP (Encapsulated Security Payload). With these protocols, IPSec ensures that transmitted data is delivered to the intended party, without augmentation or interception by unauthorized individuals. IPSec supports certificate authorities and Internet Key Exchange (IKE), and GRE is an optional configuration. IPSec encryption can be deployed in several environments and in various operating systems platforms.

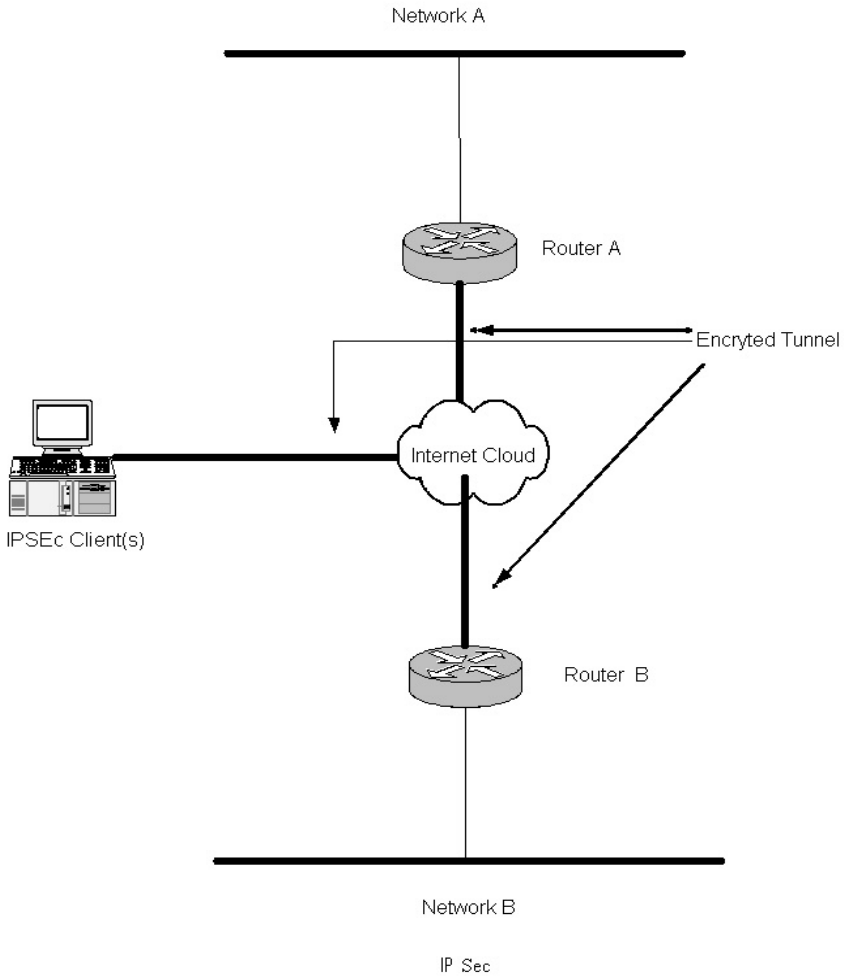


Figure 5: IPSec.

The IPSec client establishes a VPN session through the Internet with Network A and/or Network B. This session can be initiated via a direct or dialup connection. The data passing to those networks are encrypted with either 56-bit or 168-bit encryption.

7.5 Generic Routing Encapsulation Tunneling

Generic Routing Encapsulation (GRE) allows the encapsulation of IP and non-IP traffic for transmission over the Internet and/or an IP network to a specific destination. A measure of security is provided, because the packet can only enter at a specific interface. However it does not provide true security, because the packet is not encrypted.

In Fig. 6, both networks A and B run a combination of non- routable IP, IPX, and Appletalk protocols. GRE can be used to encapsulate data packets for Network A to communicate to Network B across the Internet.

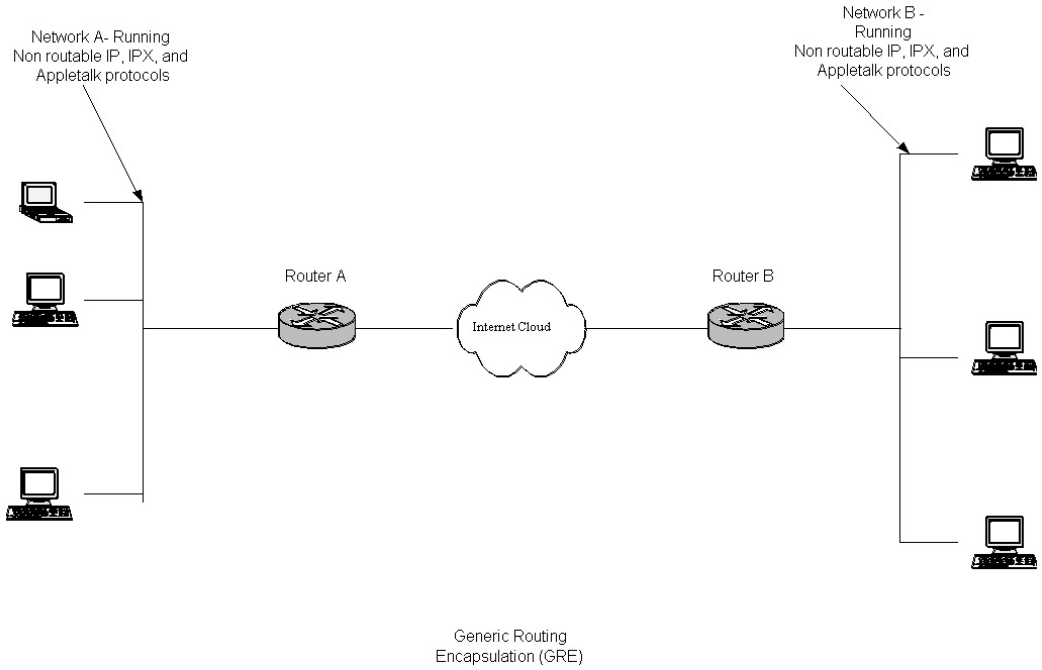


Figure 6: Generic Routing Encapsulation Tunneling.

7.6 Virtual Private Dialup Network (VPDN)

Virtual Private Dialup Network (VPDN) was developed by the Cisco Corporation and allows private network dial-in service. It may allow access to many remote servers.

The user dials into a server, often referred to as a Network Access Server (NAS; the user's destination server is known as the Home Gateway (HGW).

When a user dials into a local access server or NAS using a Point-to-Point protocol client, the NAS will forward the PPP session to the user's HGW, which will authenticate the user and initiate the session. After the user's PPP session is authenticated, then all the frames are sent through the HGW gateway router for that client. The NAS is used for access to the Internet, whereas the HGW is used for authentication to the user's home network. Authentication to NAS and HGW are not necessarily the same, but both must remain active for the remote user when she is accessing the home network.

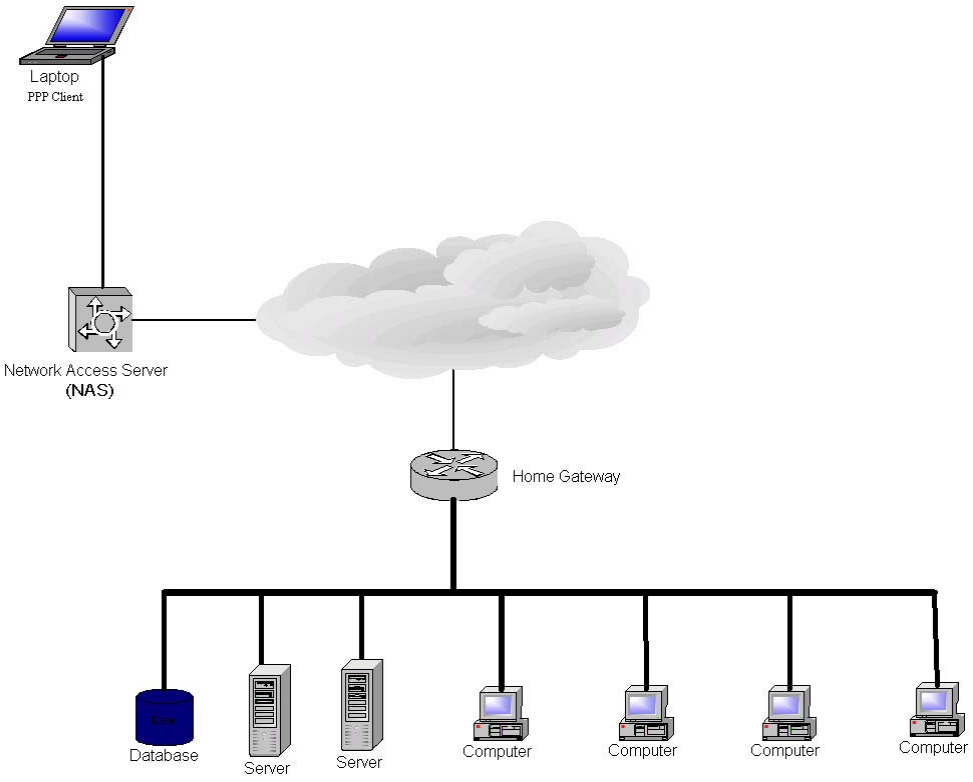


Figure 7: Virtual Private Dialup Network (VPDN).

The remote PPP client dials into the Network Access Server (NAS) using the Point-to-Point protocol (PPP). NAS will authenticate the user and pass him/her to the home gateway, which will give the user access to his/her network.

7.7 Point-to-Point Protocol over Ethernet (PPPoE) and Multiprotocol Label Switching (MPLS)

Point-to-Point Protocol over Ethernet (PPPoE) and Multiprotocol Label Switching (MPLS) are two emerging technologies. PPPoE allows Layer 3, the Network layer of Open Systems Interface (OSI), which is mainly deployed using existing Ethernet infrastructure, (e.g. DSL service), and allows multiple users access through a single access point. MPLS is an emerging technology that provides for rapid rollout and scalability.

The primary use of the VPN protocol will determine whether PPPoE or MPLS would be best for a specific application. If the main usage is to dial up and establish a VPN session, then PPTP, L2F and L2PF is the most appropriate protocol (i.e. Network-client-to-LAN connection). If the primary application is to connect to LAN or networks devices (i.e. LAN-to-LAN), then an IPSec would be the more appropriate solution. PPTP, L2F and L2FP work at the data link layer, Layer 2, of the Open Systems Interface (OSI) model. IPSec works at the Transport layer, Layer 3, of the OSI model. The advantage of working at the Data link layer is that it offers the capability to transmit non-IP traffic through tunnels. Since IPSec works at the Transport layer it is limited to using IP traffic only.

8.0 CONCLUSIONS

The key to a robust, scalable, flexible, secure, and usable network system is to establish a strong infrastructure. Consideration of all hardware systems and software applications needs to be made at the earliest possible stages. Network hardware and applications are co-dependent. It is useless to build a strong network system with built-in redundant hardware if it is not usable to the internal and external users. With the technology constantly changing it is necessary to continually review the Network Architecture. Organizations must take a proactive role in the planning and review process, which should be a standard component of good network management. Security, reliability, usability and scalability should all be a part of the normal review process.

9.0 REFERENCES

Ciolek, A. (January 4, 2001). Virtual Private Network (VPN) Security. SANS Institute. [Online]. Available: http://rr.sans.org/encryption/VPN-sec.php [1 May 2002].

Cisco Systems, Inc. (February 24, 2002). Overview of How IPSec Works. Cisco Documentation. [Online]. Available: http://www.cisco.com/univercd/cc/td/doc/product/software/ios121/121cgcr/secur_c/scprt4/scdipsec.htm#xtocid10 [5 July 2004].

Egevang, K. and Francis, P. (May 1994). Network Working Group RFC 1631: The IP Network Address Translator (NAT). Andrews System Group, Carnegie Mellon. [Online]. Available: http://asg.web.cmu.edu/rfc/rfc1631.html [5 July 2004].

Krywaniuk, A. (November 21, 2001). Security Properties of the IPSec Protocol Suite. Internet Engineering Security Taskforce. [Online]. Available: http://www.ietf.org/internet-drafts/draft-ietf-ipsec-properties-01.txt [3 May 2002].

Lee, D. (1999). Enhanced IP Services for CISCO Networks: A Practical Resource for Deploying Quality of Service, Security, IP Routing, and VPN Services. Indianapolis: Cisco Press.

USByte.com (n.d.). RAID systems. USByte.com. [Online]. Available: http://www.usbyte.com/common/raid_systems.htm [5 July 2004].

Virtual Private Network Consortium. (n.d.). VPN Standards. Virtual Private Network Consortium Homepage. [Online]. Available: http://www.vpnc.org [5 July 2004].

The Digital Library – The Bulgarian Case

Dincho Krastev
Director, Central Library
Bulgarian Academy of Sciences
1, "15 Noemvri" St., 1040 Sofia
BULGARIA

dincho@cl.bas.bg

...Teilhard de Chardin described the world as if he were outside of it. He was sure that every change, every new bifurcation, was going in the right direction – in the direction of increased spirituality. On the contrary, I am more impressed by the existence of multiple time horizons. A bifurcation can lead us to the best or to the worst. We are participating in an evolution whose outcome isn't clear to us. So I leave open the question of the meaning of being. I'm not even certain whether, put in these terms, a scientific answer is possible. Probably it has more to do with feelings or emotions. In any event, I believe it is more hopeful, more exhilarating, to be embedded in a living world than to be alone living in a dead universe. And this is really what I try to express in my work.

From Ilya Prigogine's interview, May 1983 by Robert B. Tucker (Omni Magazine)

1.0 INTRODUCTION

The analysis of the dynamics and the state of a wide interdisciplinary field such as library and information science (digital library including) presents various difficulties. It can be conducted from a number of viewpoints: phenomenological or historical, technological or sociological, futuristic or traditional, etc.; each of which is colored by different levels of optimism or pessimism of the respective researcher. Nevertheless, regardless of the scientific approach, the result of any study is usually presented in terms of analytical evaluation. This does not signify that the assessment of the changes is explicit. On the contrary, the evaluation process itself is more or less of an implicit character. In this regard, any serious analysis of the problems at hand should consciously take into account in relation to the current changes and dynamics of the social and technological realm.

Shortly after the end of the World War II the library science community began to discuss and ponder the future of the institution of the library in the context of the developments of information technology. A new "language" saw its dawn: that of information science. In general the term "information" is widely used in a range of fields without being strictly defined. This peculiar situation is probably due to the lack of a general consensus of what the notion of information signifies and what the subject of study of such an interdisciplinary area as information science is.

An overview of the subject of information science and its applications is so broad-ranging, limitless and fuzzy that in practice it stops short of being science in the narrow sense of the word. In fact, information science as a product of human culture is maybe the most telling modern example of the character of the human civilization as a continuous attempt to cope with the entropy understood in the widest sense.

According to these broad interpretations, information science deals with the generation, distribution, organization, retrieval and use of information, encoded in a range of ways, for numerous goals. Thus information is accessible through any *classical library* and its supporting information systems, and today through a variety of other media and channels such as Internet-based libraries, information portals and databases.

All these factors reveal the impact of the application of information science in the last few decades on society and demonstrate the rapid expansion of the number of ways and media which provide access to information. These developments require a re-envisioning of the institution of the library and a more modern approach to library science in which the traditional limitations regarding access to collections and records, standards and formats are surmounted.

In the last decade a more modern and encompassing view of the institution of the library and the services it provides has taken hold. According to it, the library is *everything* which preserves recorded information organized according to some criteria and to which access for the public is provided. From this standpoint a library is not only a traditional library but also any bookstore, museum, individual files and, of course, the exponentially growing Internet databases and portals.

Doubtlessly, such a comprehensive interpretation of the library reflects the real interrelations between the different subjects and institutions operating in the information market. At the same time this definition is too abstract as it puts different social institutions with widely disparate missions and goals under a common denominator. This is why I prefer to think about the library as a specific type of institution possessing a being and history of its own. Furthermore, I find it more suiting for the debate to focus not on purely scientific goals but to search for more practical and direct applications.

It is worthy remembering the slightly naive general outlook which predominated in the ‘60s and ‘70s and predicted the coming disappearance not only of paper information media but also of information intermediator to which libraries then and now happily belong. The major information clients and the individual customer need, as a rule, a very well-structured and organized information, meta-information, “knowledge”. An elementary economic analysis indicates that even the largest multinational corporations do not find it profitable on a long-term basis to invest in internal departments that provide such services. Thus, in this age the principal question to be answered in the field of library science is to what extent and how fast can existing libraries adapt to this state of things and keep its role in the sphere of information provision.

It is interesting to look into the practices and operations of a number of leading institutions and structures in the informational sphere – supranational, governmental and non-governmental, academic and commercial. From the lingo they use a variety of keywords and terms which characterize the principal information processes, and directions can be deduced. Two of these notions pop up as the most salient and characteristic together with the classical notion of a library – “digital library” and “metainformation” or “metadata”.

2.0 EXAMPLES OF BULGARIAN DIGITAL LIBRARIES

The notion of the “Digital Library” is so prominent in the media (from the project for a National Digital Library presented to the President of the US and the Senate to the priorities of the 5th and 6th Framework Programmes of the European Union) that the general public does not fully realize the significance of this development. It is natural that next to the notion of the digital library lies the concept of “metainformation” or “metadata” as any library requires structured, organized databases and information which presumes the availability of a number of indexes, secondary data and analytical information retrieved and organized on a variety of meta-levels.

I will not go into the different descriptive definitions of the “Digital Library”. I will continue relying on an implicit understanding on part of the public of this concept based on its intuitive and empirical experience.

It is quite obvious and natural that the more you go to the south of Europe or South-East of Europe, the less well-structured and organized the national library and information systems you find.

Not surprisingly, not even a single project which could be characterized as the National Digital Library Initiative has been developed in Bulgaria on a national level up until now.

So, the very few modest local initiatives with relation to what could be defined as the "Digital Library" have been started by some libraries, nongovernmental and private institutions. It is worth mentioning the following examples:

- One of the most interesting, well-developed and nicely realized virtual initiatives that I could mention is the Bulgarian WebFolk.BG initiative. It has been started by a group professional from the Bulgarian Academy of Sciences (BAS) doing research in the history of music. The leader of this initiative, Professor Lubomir Kavaldjiev, is a remarkable man with a vision of how "these things" should be done. The module they have developed is dedicated to the Bulgarian authentic folk music (http://musicart.imbm.bas.bg/default-bg.htm) and it has a lot of multimedia dimensions. What is even more interesting is that it has several levels based on users' knowledge of folk music – from the highest level (for the professionals who are quite few in numbers) to the level for the general public.
- As usual, there are a lot of full-text collections (legislations, manuals, literature). The virtual full-text library of the Bulgarian authors since ancient times to the present day, "Slovo", is one of them (http://www.slovo.bg). It has been functioning successfully for about five years, covering already quite a lot of authors and texts. It is a typical example of full-text virtual collection with an emphasis to the full-text and much less to the metadata, indexes.
- An interesting joint initiative (involving the local public library and a private company) has been started several years ago in the major city of Varna. It aimed to digitize all photo and image collections (including private) of the city. Consisting of a collection of image files with a few indexes, this unique project has been functioning successfully for a number of years (http://www.libvar.bg/old-varna/index-eng.html).
- It is worth mentioning the first online Bulgarian encyclopedia, "Trud" (http://www.encyclopedia.bg). It was developed and realized by the joint efforts of a group of people from BAS and a private publishing house.
- There are also some very nicely developed virtual art galleries, including the one developed by the Central Library of the Bulgarian Academy of Sciences (http://art.cl.bas.bg/indexcl.html). Such virtual art galleries usually consist of well-designed databases with few indexes but lack three-dimensional characteristics.
- Archeologically, as is the case for all Balkan states, Bulgaria is quite a "rich" country. Yet, apart from some technological and methodological project ideas, there has been no functioning virtual, digital model of some of the significant archeological sites. There are only Web-Based info materias and one virtual tour along the corridors of an existing museum (http://www.historymuseum.org/mainset.php3?page=2). A few years ago a virtual model of the medieval Boyana church was developed thanks to the personal efforts of a single man.
- Most of the major Bulgarian mass-media publications have their online versions. Yet these online versions are not treated strictly by the national deposit law and not a single institution is officially in charge of the preservation of these publications.

I would like now to describe the activities of the Central Library of the Bulgarian Academy of Sciences (CLBAS) with regards to digitizing Slavic manuscripts. A group of researchers led by Professor Anisava Miltenova from the Institute of Literature of BAS and CLBAS have developed a most interesting project both technologically and methodologically. The project is more closely related to the metadata than to the digital objects, images themselves. It could be titled as the "Computer Supported Processing of Archival Documents and Manuscripts and their Accessibility through Communication Technologies". Let's have a

closer look at what we call "SOFIA CORPUS OF DATA OF SLAVIC MANUSCRIPTS". I'll present here the experience of computer processing of Slavic manuscripts of CLBAS and ILBAS researchers.

We consider that an electronic database for the study of medieval manuscripts should cover three essential areas:

- Cataloging of objects (manuscripts, etc.) in an adequate structure, which contains the essential data from catalogs, e.g. signature, repository, age, material, scripture, contents, bibliographic information, etc.
- Facsimiles in the form of computerized image files, linked to the relevant entries in the catalog database. Scanning technologies available today make it possible to produce full color facsimiles of manuscripts in a satisfying quality.
- Sets of text files linked to both the relevant catalog database entries and the relevant manuscript facsimile image files. These text files should provide the monument's text, encoded according to a unified transliteration standard for further processing. Even today's sophisticated Optical Character Recognition (OCR) software packages are unable to "read" correctly manuscripts. As a result, it is still faster and economically more effective to type the text manually (in case you have the relevant drivers).

These three necessary elements include the possibility to support meta-information concerning specific media. One very effective and valuable outcome of the proposed approach of "digitization" of information for Slavic manuscripts and old printed books is that it could be processed from the available microforms, photocopies, without having the visual sources. A module which corresponds to the description of manuscripts is developed as an add-on to the detailed manuscript and old printed book description. Such a module enables the combination of partial codicological and text information, which is available only from microforms.

Another valuable component of the project is that of the bibliographical information for medieval studies. The electronic version of this part of the project began in the summer of the year 2000. Now all the relevant items on medieval Slavic languages, literature, and culture published in Bulgaria from 1990 up to the 2000 are collected and edited. The bibliography is based on the simplified Extensible Markup Language (XML) version of the Text Encoding Initiative (TEI) for bibliographic references. In this developing stage the database consists of several units including a bibliography of books, papers, and reviews, linked to cited works in each bibliographic item and to the information on the already used sources (manuscripts, old printed books, or epigraphic inscriptions).

Computer-supported research and teaching in the humanities has been growing at an increasing pace over the past decades, with new methods computer use to increase productivity in these areas. First systematic attempt to use computers in the field of Paleoslavistics took place in August 1980 at the University of Nijmegen, The Netherlands. A research team under the direction of professors A. Gruijs and C. Koster created a system for the description and cataloging of manuscripts (Producing Codicological Catalogues with the Aid of Computers). One year later, they were joined by W. Veder (Slavic Philology).

Historically, the coordination between Slavists and specialists in the fields of Latin, Greek and Hebrew paleography and codicology with respect to the medieval studies is far from being perfect. The field of mediaeval Slavonic studies used to be isolated from modern electronic tools and research and teaching methods for a long time. During the same period different hardware platforms and a wide range of software tools existed, along with the plethora of terminology and traditional topics of manuscript description used by specialists from different countries and schools.

With the Bulgarian-American project "Computer Supported Processing of Old Slavic Manuscripts" funded by IREX – Washington for the period of 1994–1995, we tried to overcome this heterogeneity of

approaches and ineffective attempts. A new type of software was built. It was based on the Standard Generalized Markup Language (SGML), accepted by the International Standards Organization (ISO), and, especially, in its TEI implementation. This undertaking was built on the framework developed within the TEI by creating a set of modifications for manuscript description.

The major template was developed in the process of the teamwork of David Birnbaum of University in Pittsburgh, USA (http://www.slavic.pitt.edu/~djb/) and Prof. Anissava Miltenova of ILBAS, Bulgaria.

The system for encoding of medieval Slavic texts (TSM) was discussed in an international conference that took place in Blagoevgrad (24th–28th July, 1995). The reports from the conference were published in a separate volume (Birnbaum, Boyadzhiev, Dobreva, Miltenova 1995). The philosophy of SGML helped to settle some well-known misunderstandings among paleoslavists concerning philological questions of terminology, inventory of units, character sets and data structure.

At this point the group of Prof. Anissava Miltenova has followed five main principles, formulated by David J. Birnbaum: 1) Standardization of document file format; 2) Multiple use (ensured by the separation of data from processing); 3) Portability of electronic texts (independence of local platforms); 4) Necessity of long-term preservation of manuscripts and archival documents in electronic form; and 5) Orientation towards well-structured divisions of data according to established traditions of codicology, textology, paleography, etc.

The movement from a relational database management system (RDBMS) framework to SGML marked a significant reorientation in the conceptualization of computer-assisted manuscript description. More importantly, though, our SGML-based undertaking was oriented towards preparing manuscript descriptions that might be suitable for printing, electronic rendering, and searching, as was the case with the RDBMS approach.

We can observe, though, that attempts to describe Slavic manuscripts in electronic form prior to 1994 relied almost exclusively on relational or flat-file databases, an architecture that is well suited to the record-and-field nature of some bibliographic information, but that is poorly designed for representing the hierarchical structures and blocks of prose that are more natural in manuscript description.

We anticipated even at that stage (prior to 1994) that the manuscript description files would be suitable for direct analysis, so that we would be able, for example, to identify patterns of structural similarity within a corpus of manuscripts on the basis of the same raw data files that we would also use to generate traditional printed manuscript descriptions.

This database development represented important first steps in the conceptualization of Slavic manuscript description as a problem of information science, and not merely of descriptive philology and codicology, but architectural limitations inherent in the RDBMS architecture prevented these undertakings from exercising any significant, long-term influence on the practice of Slavic manuscript studies

Within the framework of the first **pilot** (experimental) project, over **three hundred fifty manuscripts** were processed by using TSM system in the SGML environment with the corresponding interface A/E (*Author/Editor,* SoftQuad, Canada) software package. Scientific papers and indices of the pilot project were published under the title *Medieval Slavic manuscripts and SGML: Problems and perspectives* (2000).

We consider our prior close collaboration with specialists in Slavic and general humanities computing (e.g., Institute for Computational Linguistics, Pisa, Italy; and Portsmouth University, Great Britain) to be one of the strongest features of our both evaluative feedback on our proposals and means for ensuring that our results will reach authoritative figures and institutions. A new stage of the project was the joint work with Prof. Ralph Cleminson on cataloging of early printed books in Great Britain and with

Dr. Martha Boyanivska (Ukraine) on description of Slavic manuscripts in the collection of National Museum in Lviv. Last year (2003) a joint contract was signed between the British Library and the CLBAS, having as a major target the processing by our group of professionals a certain collection of Slavic manuscripts from the BL. Recently a similar contract was signed with the library of the Russian Academy of Sciences in St Petersburg There is a strong interest in starting such joint projects with the National Library of St Petersburg, Russia and the State Library in Odessa, Ukraine. The same type of project is going on with Sweden (official partner Royal Academy of Sciences and as sub-partners all major Swedish libraries with Slavic manuscript collections).

The Sofia project activities nowadays are concentrated on the following main fields:

- The first of these is the development of the model for the processing of specifically Slavonic manuscripts and the provision, in an adequate structure, of data fields for the cataloging of manuscripts.
- Next is the use of these principles and software to produce a database of descriptions of manuscripts in Bulgaria and, ultimately, elsewhere (also an "electronic catalog").
- The descriptions of the manuscripts themselves constitute the first of these elements. The second will contain facsimiles in the form of computerized picture files, linked to the relevant entries in the catalog database.
- Quite an important field is the development of auxiliary materials and databases ("electronic reference books") for the study of Slavonic manuscripts, in many cases by extrapolation of the data assembled in the other phases of the project. Part of this field consists of bibliographic database for the described sources.
- As a necessary part of the manuscript description the model for digitization of microforms is developed.

These ideas have been discussed at a special panel in the framework of the 12th International Congress of Slavists, Krakow, 1998. Participants from Byelorussia, Bulgaria, Czech Republic, Finland, Italy, Macedonia, Great Britain, the US, etc. put on discussion some mainstream questions in the field. One of the results from this discussion was the establishment of a Commission to the Executive Council of the Congress for Computer Supported Processing of Slavic Manuscripts and Early Printed Books.

Part of these activities is also the Master Program at the Faculty of Slavic Studies at the University of Sofia that has been started. Within the framework of it an essential attention is given to the knowledge in the fields of markup languages, electronic transcription, text corpuses and text analysis. The Master Program also includes student training in computational linguistics and students' own work on implementation of computer tools in humanities (www.slav.uni-sofia.bg/Pages/comhuen.htm).

The other principal achievement of this time was the development by Stanimir Velev of a query interface for the manuscript descriptions that was prepared within the Repertorium project. Its interface was an interim solution that has now been superseded by Extensible Stylesheet Language for Transformations (XSLT) scripting, but for several years it served as the principal query engine for scholars at the Institute of Literature who were conducting philological research on the basis of our manuscript descriptions.

Our collection of articles was the first demonstration of the utility of SGML files in traditional philological research, although at that time the principal type of processing involved structured searching, a very powerful feature of SGML, but one that only scratches the proverbial surface of the capabilities of such a system.

The last phase of this process is characterized not only by the accumulation of still more manuscript descriptions, but also by the conversion of our materials from SGML to XML. The transition to XML was

dictated by the remarkably broad acceptance of XML within the electronic-text community, and particularly by its adoption by the TEI, initially as an alternative to SGML, but ultimately as a replacement for it. We have currently converted over one hundred manuscript descriptions from our initial corpus of three hundred; the rest will be converted in time, and all new descriptions are being created directly in XML.

While XML was attractive because of its status as an emerging standard with a very wide following, it was also appealing because of the ancillary standards that were developed in coordination with it. In particular, we found XSLT particularly well-suited to processing XML manuscript descriptions in order to generate different views of the data, and it also provided a standards-based alternative to the database orientation that was implemented. We also used SVG, Scalable Vector Graphics, to generic graphic representations of manuscript structures. As Tommie Usdin said at the Extreme Markup 2003 conference in Montreal, "XML has made true all of the lies we told about SGML." By this she meant that SGML promised structured descriptions that could be transformed and visualized in new and different ways, but in the early days of SGML, one had to encode manuscript descriptions while taking on faith that the transformation and visualization tools would eventually be developed. XSLT and SVG have made it possible for XML to deliver the transformations and visualizations that were only foreseen, but not actually, within the early SGML context.

3.0. CONCLUSION

I would like to emphasize that, after using portable electronic files in SGML/XML format, several scientists have changed their point of view on the effectiveness of the applications of modern software tools to manuscripts and medieval texts. It is obvious how deep into the structure of medieval texts nowadays a researcher could go. Computer and software tools that are in use for the creation and maintenance of the Sofia database at the beginning of the 21st century are very powerful research instruments, more accurate and more comfortable for the users than they were only a few years ago. Using SGML/XML-like encoding guarantees compatibility, interchange, and multiple uses of electronic editions – which is very important both for research work and for preservation of manuscripts in the libraries. We need to continue the team work, because it is the only possible organization of such kind of projects. Especially important are the efforts of more libraries and archives to be involved as a common unified effort in order to preserve and make more accessible these most valuable medieval manuscripts and archival documents. Of course, a strong international cooperation and exchange of information in the field of computational medieval studies and computational humanities in general is also essential today and even more for the future.

4.0 REFERENCES AND FURTHER READING

Birnbaum, D.J., Boyadzhiev, A.T., Dobreva, M., and Miltenova, A.L. (eds.) (1995). *Computer Processing of Medieval Slavic Manuscripts. Proceedings.* First International Conference, 24-28 July 1995, Blagoevgrad, Bulgaria. Sofia: Marin Drinov Publishing House.

Computational Approaches to the Study of Early and Modern Slavic Languages and Texts. (2003). Ed. by David Birnbaum, Anissava Miltenova, and Sarah Slevinski. Sofia.

Medieval Slavic Manuscripts and SGML: Problems and Perspectives. (2000). Ed. by Anissava Miltenova and David Birnbaum. Sofia.

Scripta & e-Scripta. The Journal of Interdisciplinary Medieval Studies. 1, 2003. Ed. by Anissava Miltenova.

NORTH ATLANTIC TREATY ORGANISATION

RESEARCH AND TECHNOLOGY ORGANISATION

BP 25
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int

DIFFUSION DES PUBLICATIONS

RTO NON CLASSIFIEES

Les publications de l'AGARD et de la RTO peuvent parfois être obtenues auprès des centres nationaux de distribution indiqués ci-dessous. Si vous souhaitez recevoir toutes les publications de la RTO, ou simplement celles qui concernent certains Panels, vous pouvez demander d'être inclus soit à titre personnel, soit au nom de votre organisation, sur la liste d'envoi.

Les publications de la RTO et de l'AGARD sont également en vente auprès des agences de vente indiquées ci-dessous.

Les demandes de documents RTO ou AGARD doivent comporter la dénomination « RTO » ou « AGARD » selon le cas, suivi du numéro de série. Des informations analogues, telles que le titre est la date de publication sont souhaitables.

Si vous souhaitez recevoir une notification électronique de la disponibilité des rapports de la RTO au fur et à mesure de leur publication, vous pouvez consulter notre site Web (www.rta.nato.int) et vous abonner à ce service.

CENTRES DE DIFFUSION NATIONAUX

ALLEMAGNE
Streitkräfteamt / Abteilung III
Fachinformationszentrum der
Bundeswehr (FIZBw)
Friedrich-Ebert-Allee 34, D-53113 Bonn

BELGIQUE
Etat-Major de la Défense
Département d'Etat-Major Stratégie
ACOS-STRAT – Coord. RTO
Quartier Reine Elisabeth
Rue d'Evère, B-1140 Bruxelles

CANADA
DSIGRD2
Bibliothécaire des ressources du savoir
R et D pour la défense Canada
Ministère de la Défense nationale
305, rue Rideau, 9e étage
Ottawa, Ontario K1A 0K2

DANEMARK
Danish Defence Research Establishment
Ryvangs Allé 1, P.O. Box 2715
DK-2100 Copenhagen Ø

ESPAGNE
SDG TECEN / DGAM
C/ Arturo Soria 289
Madrid 28033

ETATS-UNIS
NASA Center for AeroSpace
Information (CASI)
Parkway Center, 7121 Standard Drive
Hanover, MD 21076-1320

FRANCE
O.N.E.R.A. (ISP)
29, Avenue de la Division Leclerc
BP 72, 92322 Châtillon Cedex

GRECE (Correspondant)
Defence Industry & Research
General Directorate, Research Directorate
Fakinos Base Camp, S.T.G. 1020
Holargos, Athens

HONGRIE
Department for Scientific Analysis
Institute of Military Technology
Ministry of Defence
H-1525 Budapest P O Box 26

ISLANDE
Director of Aviation
c/o Flugrad
Reykjavik

ITALIE
Centro di Documentazione
Tecnico-Scientifica della Difesa
Via XX Settembre 123
00187 Roma

LUXEMBOURG
Voir Belgique

NORVEGE
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

PAYS-BAS
Royal Netherlands Military
Academy Library
P.O. Box 90.002
4800 PA Breda

POLOGNE
Armament Policy Department
218 Niepodleglosci Av.
00-911 Warsaw

PORTUGAL
Estado Maior da Força Aérea
SDFA – Centro de Documentação
Alfragide
P-2720 Amadora

REPUBLIQUE TCHEQUE
DIC Czech Republic – NATO RTO
LOM PRAHA s. p.
o.z. VTÚL a PVO
Mladoboleslavská 944, PO BOX 16
197 21 Praha 97

ROYAUME-UNI
Dstl Knowledge Services
Information Centre, Building 247
Dstl Porton Down
Salisbury
Wiltshire SP4 0JQ

TURQUIE
Milli Savunma Bakanlığı (MSB)
ARGE ve Teknoloji Dairesi Başkanlığı
06650 Bakanliklar – Ankara

AGENCES DE VENTE

NASA Center for AeroSpace Information (CASI)
Parkway Center, 7121 Standard Drive
Hanover, MD 21076-1320
ETATS-UNIS

The British Library Document Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
ROYAUME-UNI

Canada Institute for Scientific and Technical Information (CISTI)
National Research Council
Acquisitions, Montreal Road, Building M-55
Ottawa K1A 0S2, CANADA

Les demandes de documents RTO ou AGARD doivent comporter la dénomination « RTO » ou « AGARD » selon le cas, suivie du numéro de série (par exemple AGARD-AG-315). Des informations analogues, telles que le titre et la date de publication sont souhaitables. Des références bibliographiques complètes ainsi que des résumés des publications RTO et AGARD figurent dans les journaux suivants :

Scientific and Technical Aerospace Reports (STAR)
STAR peut être consulté en ligne au localisateur de ressources uniformes (URL) suivant:
http://www.sti.nasa.gov/Pubs/star/Star.html
STAR est édité par CASI dans le cadre du programme NASA d'information scientifique et technique (STI)
STI Program Office, MS 157A
NASA Langley Research Center
Hampton, Virginia 23681-0001
ETATS-UNIS

Government Reports Announcements & Index (GRA&I)
publié par le National Technical Information Service
Springfield
Virginia 2216
ETATS-UNIS
(accessible également en mode interactif dans la base de données bibliographiques en ligne du NTIS, et sur CD-ROM)

NORTH ATLANTIC TREATY ORGANISATION

BP 25
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int

RESEARCH AND TECHNOLOGY ORGANISATION

DISTRIBUTION OF UNCLASSIFIED RTO PUBLICATIONS

AGARD & RTO publications are sometimes available from the National Distribution Centres listed below. If you wish to receive all RTO reports, or just those relating to one or more specific RTO Panels, they may be willing to include you (or your Organisation) in their distribution.

RTO and AGARD reports may also be purchased from the Sales Agencies listed below.

Requests for RTO or AGARD documents should include the word 'RTO' or 'AGARD', as appropriate, followed by the serial number. Collateral information such as title and publication date is desirable.

If you wish to receive electronic notification of RTO reports as they are published, please visit our website (www.rta.nato.int) from where you can register for this service.

NATIONAL DISTRIBUTION CENTRES

BELGIUM
Etat-Major de la Défense
Département d'Etat-Major Stratégie
ACOS-STRAT – Coord. RTO
Quartier Reine Elisabeth
Rue d'Evère
B-1140 Bruxelles

CANADA
DRDKIM2
Knowledge Resources Librarian
Defence R&D Canada
Department of National Defence
305 Rideau Street
9th Floor
Ottawa, Ontario K1A 0K2

CZECH REPUBLIC
DIC Czech Republic – NATO RTO
LOM PRAHA s. p.
o.z. VTÚL a PVO
Mladoboleslavská 944, PO BOX 16
197 21 Praha 97

DENMARK
Danish Defence Research Establishment
Ryvangs Allé 1
P.O. Box 2715
DK-2100 Copenhagen Ø

FRANCE
O.N.E.R.A. (ISP)
29, Avenue de la Division Leclerc
BP 72
92322 Châtillon Cedex

GERMANY
Streitkräfteamt / Abteilung III
Fachinformationszentrum der Bundeswehr (FIZBw)
Friedrich-Ebert-Allee 34
D-53113 Bonn

GREECE (Point of Contact)
Defence Industry & Research General Directorate, Research Directorate
Fakinos Base Camp, S.T.G. 1020
Holargos, Athens

HUNGARY
Department for Scientific Analysis
Institute of Military Technology
Ministry of Defence
H-1525 Budapest P O Box 26

ICELAND
Director of Aviation
c/o Flugrad, Reykjavik

ITALY
Centro di Documentazione Tecnico-Scientifica della Difesa
Via XX Settembre 123
00187 Roma

LUXEMBOURG
See Belgium

NETHERLANDS
Royal Netherlands Military Academy Library
P.O. Box 90.002
4800 PA Breda

NORWAY
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

POLAND
Armament Policy Department
218 Niepodleglosci Av.
00-911 Warsaw

PORTUGAL
Estado Maior da Força Aérea
SDFA – Centro de Documentação
Alfragide, P-2720 Amadora

SPAIN
SDG TECEN / DGAM
C/ Arturo Soria 289
Madrid 28033

TURKEY
Milli Savunma Bakanlığı (MSB)
ARGE ve Teknoloji Dairesi Başkanlığı
06650 Bakanliklar – Ankara

UNITED KINGDOM
Dstl Knowledge Services
Information Centre, Building 247
Dstl Porton Down
Salisbury, Wiltshire SP4 0JQ

UNITED STATES
NASA Center for AeroSpace Information (CASI)
Parkway Center, 7121 Standard Drive
Hanover, MD 21076-1320

SALES AGENCIES

NASA Center for AeroSpace Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320
UNITED STATES

The British Library Document Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
UNITED KINGDOM

Canada Institute for Scientific and Technical Information (CISTI)
National Research Council
Acquisitions
Montreal Road, Building M-55
Ottawa K1A 0S2, CANADA

Requests for RTO or AGARD documents should include the word 'RTO' or 'AGARD', as appropriate, followed by the serial number (for example AGARD-AG-315). Collateral information such as title and publication date is desirable. Full bibliographical references and abstracts of RTO and AGARD publications are given in the following journals:

Scientific and Technical Aerospace Reports (STAR)
STAR is available on-line at the following uniform resource locator:
http://www.sti.nasa.gov/Pubs/star/Star.html
STAR is published by CASI for the NASA Scientific and Technical Information (STI) Program
STI Program Office, MS 157A
NASA Langley Research Center
Hampton, Virginia 23681-0001
UNITED STATES

Government Reports Announcements & Index (GRA&I)
published by the National Technical Information Service
Springfield
Virginia 2216
UNITED STATES
(also available online in the NTIS Bibliographic Database or on CD-ROM)