OECD

STI

REVIEW No. 24

SCIENCE TECHNOLOGY INDUSTRY

Special Issue on "The Global Research Village"

Maximising the Benefits of Information Technology for Science: Overview and Major Issues

A Global Research Infrastructure for the Global Research Village: The European Perspective

Computer Networks and the Virtual College

Electronic Publishing Issues

Denmark's Electronic Research Library: A Tool for Institutional Change

The Future of Mathematical Databases

Information Infrastructures in the Social Sciences

© OECD, 1999.

© Software: 1987-1996, Acrobat is a trademark of ADOBE.

All rights reserved. OECD grants you the right to use one copy of this Program for your personal use only. Unauthorised reproduction, lending, hiring, transmission or distribution of any data or software is prohibited. You must treat the Program and associated materials and any elements thereof like any other copyrighted material.

All requests should be made to:

Head of Publications Service,
OECD Publications Service,
2, rue André-Pascal, 75775 Paris
Cedex 16, France.

No. 24

STI
REVIEW

Special Issue on
“The Global Research Village’’

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Pursuant to Article 1 of the Convention signed in Paris on 14th December 1960, and which came into force on 30th September 1961, the Organisation for Economic Co-operation and Development (OECD) shall promote policies designed:

– to achieve the highest sustainable economic growth and employment and a rising standard of living in Member countries, while maintaining financial stability, and thus to contribute to the development of the world economy;

– to contribute to sound economic expansion in Member as well as non-member countries in the process of economic development; and

– to contribute to the expansion of world trade on a multilateral, non-discriminatory basis in accordance with international obligations.

The original Member countries of the OECD are Austria, Belgium, Canada, Denmark, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States. The following countries became Members subsequently through accession at the dates indicated hereafter: Japan (28th April 1964), Finland (28th January 1969), Australia (7th June 1971), New Zealand (29th May 1973), Mexico (18th May 1994), the Czech Republic (21st December 1995), Hungary (7th May 1996), Poland (22nd November 1996) and Korea (12th December 1996). The Commission of the European Communities takes part in the work of the OECD (Article 13 of the OECD Convention).

Également publié en français

STI REVUE
Numéro spécial :
« Le village mondial de la recherche »
N° 24

© OECD 1999
Permission to reproduce a portion of this work for non-commercial purposes or classroom use should be obtained through the Centre français d’exploitation du droit de copie (CFC), 20, rue des Grands-Augustins, 75006 Paris, France, Tel. (33-1) 44 07 47 70, Fax (33-1) 46 34 67 19, for every country except the United States. In the United States permission should be obtained through the Copyright Clearance Center, Customer Service, (508)750-8400, 222 Rosewood Drive, Danvers, MA 01923 USA, or CCC Online: http://www.copyright.com/. All other applications for permission to reproduce or translate all or part of this book should be made to OECD Publications, 2, rue André-Pascal, 75775 Paris Cedex 16, France.

FOREWORD

Prepared by the OECD Directorate for Science, Technology and Industry, the *STI Review*, published twice yearly, presents studies of interest to science, technology and industry policy makers and analysts, with particular emphasis on cross-country comparisons, quantitative descriptions of new trends and identification of recent and future policy problems. Because of the nature of OECD work, the *STI Review* explores structural and institutional change at global level as well as at regional, national and sub-national levels. Issues often focus on particular themes, such as surveys of firm-level innovation behaviour and technology-related employment problems.

This issue of the *STI Review* focuses on the impacts of information and communication technologies on scientific research. The OECD Committee for Scientific and Technological Policy is examining the changing role of governments in supporting the science system, a change triggered to a large extent by the significant potential that these technologies offer. The papers in this issue are drawn mainly from the second Conference on the Global Research Village on maximising the benefits of information technology for science, held in Sintra, Portugal on 17-18 September 1998 and organised jointly by the Portuguese Ministry for Science and Technology and the OECD. The main themes of the papers are the development of ICT-based infrastructure for science in different parts of the OECD and the use of this infrastructure by scientists for communicating, collaborating, accessing information and publishing research results.

The views expressed in this publication do not necessarily reflect those of the Organisation or of its Member countries. The *STI Review* is published on the responsibility of the Secretary-General of the OECD.

TABLE OF CONTENTS

MAXIMISING THE BENEFITS OF INFORMATION TECHNOLOGY FOR SCIENCE: OVERVIEW AND MAJOR ISSUES

Introduction

Information and communication technologies (ICT) have become essential support tools for the activities carried out by researchers. Extraordinary advances in these technologies have influenced almost every aspect of scientific activity, from data collection and analysis, through ways of linking scientists around the world with new collaborative structures, to means of disseminating research outputs. Changes induced by the use of ICT have come at a time when science systems were already facing the need to adapt to other changes – such as falling government support for R&D, growing emphasis on economic relevance and accountability, closer links with society and the business sector, greater attention to diffusion and international co-operation in research – and simultaneously assuming greater importance in increasingly knowledge-based economies.

Many governments are consequently giving increasing attention to the role of ICT in the re-definition of individual and collective government responsibilities towards the science system. In the context of the OECD, a conference on the Global Research Village, convened by the Committee for Scientific and Technological Policy and held in Denmark in June 1996, initiated an ongoing dialogue between policy makers and the scientific community to determine an appropriate role for governments in facilitating the provision and productive use of ICT resources in the global science arena.

A report prepared by the OECD Secretariat during 1997 then examined how the use of ICT influenced science in the five following areas: communication among scientists; access to scientific information; scientific instruments; electronic publishing in science; and science education and training, considering both benefits and implementation issues. On the basis of this background report, a second conference on the Global Research Village, held in Sintra in September 1998, focused on maximising the potential benefits of ICT for science. More than 60 participants, consisting of senior government officials from Ministries of Science, Technology, Industry and Education and from intergovernmental organisations, as well as representatives of research institutes, academia and industry, discussed the role of government in realising the potential of ICT for science.

This paper first highlights differences in ICT infrastructure development in various parts of the OECD. It then summarises the impacts of ICT on certain aspects of the science system, that is, on communication among scientists, on scientific publishing, on research libraries and on scientific databases. Finally, it presents the general policy recommendations formulated at the Sintra Conference.

ICT infrastructure issues

A wide range of developments in ICT covering hardware, software and networking technologies underlie ongoing changes in the science system. The physical ICT infrastructure constitutes the foundation of ICT-based applications that not only support scientists in their daily tasks but also allow them to carry out research projects that would have been inconceivable in the past. There is an intense need for an adequate physical ICT infrastructure to support activities such as data acquisition, transmission and sharing, computational research, access to scientific instruments, informal communications among scientists and formal dissemination of research outputs. The growth of the Internet and new interactive, often multimedia, applications that require the transmission of large volumes of data has led to rapidly growing demand for high bandwidth and performance.

United States

ICT-related infrastructure is particularly well developed in the United States. The Internet, initially developed for defence purposes, went through different stages of funding and management until 1995 when further investment was left in the hands of increasingly active commercial interests. The government nevertheless continues to be very active in promoting advanced research with the aim of pushing the capability of supercomputers, networks, storage devices and software beyond their current limits in the interest of science users (*Bainbridge*).

In the area of networking and applications, the NSF has been developing the very High Speed Backbone Network Service (vBNS) to upgrade the existing Internet's backbones, the primary hubs of data transmission, initially to 622 megabits per second (Mbps) and then to 2.4 gigabits per second (Gbps). A co-operative project by universities, Internet2 aims to extend the capacity of upgraded backbones to the campuses of participating universities (initial transmission speed of 2.4 Gbps between core nodes by January 1999, later increasing to 9.6 Gbps) and to develop software for the new network. Both projects involve partnership with industry. In addition, the government supports these efforts by focusing on advanced applications with the Next Generation Internet (NGI) Initiative (to date, funds have been used to connect universities to vBNS and to improve services on high-speed computer networks).

As government continues to be the principal customer for the kind of computing capacity offered by supercomputers (for modelling and simulating complex natural phenomena), government investment continues to be essential for further advances in this area in spite of very rapid ICT development by industry. Moreover, the government is continuing substantial investments in fundamental software research, an area where industry has invested much less compared to its investments in hardware. At the same time, requirements related to collecting, storing, analysing and sharing data have gained importance in this country and are receiving more attention.

European Union

The situation is quite different in Europe, where a pervasive high-performance network infrastructure is lacking (*Axmann and Payr*). Member states of the European Union have each funded the development of National Research and Education Networks (NRENs) functioning at varying transmission speeds and generally open to a wide constituency of researchers, teachers and students in a given country. It has, nevertheless, proven to be very hard to build trans-European networks, as organising and funding a European backbone infrastructure has been hampered by the existence of national telecommunications monopolies until the recent past. Bandwidth between French and Belgian academic communities was as low as 6 Mbps in 1998, for example.

Regulatory reform of the telecommunications sector, where high costs and prices have hampered the use of ICT, has been proceeding in Europe, with complete deregulation in most countries since January 1998. However, more remains to be done especially in some European countries which are slow to introduce new market structures in spite of efforts by the European Commission to actively pursue deregulation/liberalisation of the telecommunications sector. The industry is still in the process of restructuring and has not been active in research networking.

The only possibility so far to promote research networking at the European level has been through European Community programmes for R&D. The Fourth Framework Programme which could only support research in networking and not research networks, nevertheless led to the creation in 1997 of TEN-34, a network nominally running at 34 Mbps that will be upgraded to 155 Mbps and that will allow NRENs to connect at various speeds. It is only in the Fifth Framework Programme (1998-2002) that direct funding of networks for research could be achieved, specifically through joint funding by the EC (40%) and by member states (60%), amounting to US$ 172 million.

Asia-Pacific region[1]

Efforts to connect research and education establishments in this region are relatively recent but are moving at a steady pace. The Asia-Pacific Advanced Network (APAN), linking Australia, China, Hong Kong (China), Indonesia, Japan, Korea, Malaysia, the Philippines, Singapore, Thailand and the United States, was launched in 1997 to establish high-speed links – between 45 and 155 Mbps – for research and education among participating countries. This network is funded by the National Science Foundation of the United States and by the Science and Technology Agency of Japan. Its aim is to alleviate regional disparities in Internet access. There are plans to extend this network to Europe and Latin America in 1999.

Nevertheless, access to the Internet is still insufficient for scientists in some countries of the region, such as Cambodia, Laos and Vietnam, to name but a few. The Asia Pacific Networking Group was formed to co-ordinate efforts in this area and in the education and training of Internet engineers. Language is also a barrier in non-English speaking parts of the Asia-Pacific region, both in terms of communications between scientists and in terms of the time it can take to adapt new software tools – two years in the case of Adobe PDF tools. Graphic communication tools have consequently gained importance here compared to other regions as advances in technology facilitate the transmission of still or moving images.

In Japan, in August 1995 the ministries and agencies concerned with the research information infrastructure developed guidelines on the implementation of information technology in research and development. Ongoing efforts aim to strengthen the research information infrastructure by linking research organisations with a network that transcends boundaries between research institutes, government ministries and agencies as well as between countries. The Science and Technology Agency has a five-year, US$ 400-million project to expand IMnet, the interministry research information network which is operated by the Japan Science and Technology Corporation since 1997 and which is part of APAN. The objective is to develop a domestic backbone with a transmission capacity of 1 Gbps. Efforts are also directed at improving content with the expansion and maintenance of a research information database and of factual databases.

The Ministry of Education, Science, Sports and Culture has been establishing and upgrading scientific information distribution systems such as networks and databases. Japan's Science Information Network (SINET) provided by the National Centre for Science Information System (NACSIS) in the Ministry of Education links universities and other institutions nation-wide. The Ministry is speeding up the network, currently operating at speeds from 6 to 150 Mbps and extending its international links, which currently cover Thailand, the United States and the United Kingdom. The link to the United States, which has particularly heavy traffic, was upgraded to 45 Mbps in October 1997. As of March 1997, the network linked

613 institutions. The NACSIS also maintains databases on catalogues of all the Japanese universities, lists of Japanese researchers and abstracts of many scientific papers. Japanese and English versions of the catalogues can be accessed through the World Wide Web.

Diverse stages of development and related issues

Given that the research infrastructure is at very different stages of development in different parts of the OECD and is currently far from global, the issues raised vary by country and region. In the United States, where the basic research network has already been substantially developed, government emphasis has moved on to stimulating connections by universities to the backbone network and to advanced research in various components of the infrastructure that will enable scientists to push back the frontiers of science. Parallel investments in hardware, software modules, databases and institutions for training researchers should help to maximise the payoff from the physical infrastructure.

In the European Union, where separate national research networks connected by internetworking services provided by national public telecommunications operators have led to an inadequate research infrastructure with organisational and funding difficulties, the emphasis today is on achieving liberalisation of the telecommunication industry and on creating the adequate trans-European backbone infrastructure. The new model for financing ICT infrastructure of the Fifth Framework Programme for R&D is expected to prove more efficient than the earlier one, although budgetary uncertainty continues. In addition, there are calls for a different organisational structure and for an independent operator for the European research network. Although the US situation is frequently cited as the model which Europe should follow in the area of connections, the underlying philosophy in Europe has been to place greater emphasis on breadth of access by all research and education institutions to a knowledge network, rather on than leading-edge technology and advanced applications.

At the global level, further efforts will be needed if there is a real intention to integrate scientific communities worldwide. Countries participating in APAN, cognisant of the widening gap between scientists in the Asia-Pacific region who have Internet access and those who do not, are already seeking funding from public and private sources to alleviate these disparities. In Europe, most efforts today focus on linking member states of the European Union and connections with and between other countries are lagging.

Demand by all countries for connections to the United States has been particularly strong in the past, with the consequence that these countries have had to bear the lion's share of the investments needed to establish links with the United States' networks and institutions. Now that demand and the flow of traffic are more evenly distributed into and out of the United States, there has been mounting pressure for equity in cost and tariff sharing between the United States and other countries. New arrangements regarding trans-continental network funding, equitable sharing of costs and pricing now need to be made.

Government support continues to be needed to foster the development and availability of the technological infrastructure that underlies the work of scientists and to ensure that scientists are trained to make efficient use of the ICT infrastructure. Leading-edge technologies that are beyond the scope of commercial interest, broadly those that enable researchers to work in distributed co-operative environments and particularly those for the remote control of instruments, cannot be developed without government support. Government involvement will continue to be necessary even if commercial interests were to make the investments required for advanced scientific purposes, since funding the use of networks by scientists would then become an issue. Governments must take ICT needs specifically into consideration when funding science. More generally, for the science networks to develop as did the telephone networks – with the help of the market – isolation of the science establishment and its networks must be avoided and an expansion benefiting a broader segment of society must be encouraged as far as possible.

Impacts of ICT on research practices and productivity

One of the more mature applications of ICT involves the use of the Internet for informal communications among scientists. Apart from greatly enhancing the quantity, quality and speed of communication among scientists, ICT use has also had various effects on the organisation of work in science (*Walsh*). However, the overall impacts of new research practices on scientific outcomes are not yet clear.

Evolving research practices

Some of the effects of ICT-based communications are quite clear. Improved communication through ICT has contributed to an increase in the size of professional networks and has led to an increase in remote collaboration, particularly at international level. It has enabled researchers to overcome many barriers to communication due to geographic distance, such as time, cost and language. ICT has not only facilitated communications with colleagues but has also provided access to equipment, software and databases.

With tighter links among geographically dispersed scientists, the international community of scholars has become more dense. For a given research topic, ICT has allowed the creation of more complex work groups with more fluid structures where research topic, rather than geographic proximity, determines collaboration decisions.

The frequency and amount of communication during a research project has increased. This is particularly important in the context of long-term experiments, shift work, and different time zones since it may increase attachment to the research group, job satisfaction and commitment, and ensure that various elements of lengthy research projects remain well co-ordinated. There is also evidence that e-mail communication does away with much of the socialising that generally accompanies face-to-face or even telephone conversations. This may result in less collegiality and a more alienating work environment, but may also lead to increased focus on business matters.

There is conflicting evidence on the net effect of some other changes induced by ICT, however, notably on status and hierarchy. While ICT can facilitate access by younger scientists and scientists at peripheral institutions to crucial remote resources such as colleagues, and information as well as computing facilities, software or databases that have traditionally been unequally distributed, it is not clear whether they can actively participate in projects or make significant contributions. Another uncertainty relates to whether ICT leads to greater cross-disciplinary collaboration, or to the fragmentation of research with researchers using limited communication time to interact only with those in their speciality. While this could be beneficial and achieve economies of scale in certain scientific areas, it could also reduce cross-fertilisation of ideas among disciplines.

The net impact of these new forms of organisation is not yet clear. Overall, ICT use may lead more to a broadening of the science base than to a change in the hierarchy of scientific institutions. In addition to aspects on which there is conflicting evidence, uncertainty arises from the fact that impacts are tempered by use which varies significantly by field, as does the potential to benefit from the technology. Fields in which interdependence is high, with frequent interaction between collaborators, as well those in which collaborators tend to be dispersed – such as mathematics, physics and aerospace – are most likely to benefit. In contrast, in fields such as ornithology and botany, technical limitations related to the transmission of non-textual information and a relatively slow pace of discovery may limit benefits.

Productivity

Enhanced communications and co-operation among scientists is but one aspect of ICT-induced developments contributing to the ongoing transformation of the science system. Others, such as greater and faster access to scientific information, more rapid and wider dissemination of research results, increasingly powerful, effective and accessible scientific instruments, and improved and more accessible digital resources for the education and training of scientists are all contributing to this transformation.

Nevertheless, whether these developments are helping to lower the costs of research or improve the productivity of the science system remains unclear. Evidence concerning the use of networks and the productivity of scientists is correlational rather than causal. The productivity of scientists, when measured by the number of articles produced, totally ignores the more important quality aspect and needs to be weighted, for example, by considering the extent of citations. In addition, although ICT seem to improve the efficiency of scientists, both the extent of use and the potential benefits vary by scientific discipline. Furthermore, the use of ICT in science may involve learning costs that reduce the potential gains in productivity.

ICT may also lead to economies of scale and scope, but collaborative arrangements are increasingly specialised while an increasing number of scientific breakthroughs cross disciplinary boundaries. Electronic communication is more useful for transmitting codified knowledge and less so for the diffusion of know-how or skills embodied in people, considered to be increasingly relevant in OECD economies. Overall, leading-edge scientific advances remain expensive and in some cases the costs may be increasing. In times of tight budgets, increases in productivity, if any, may contribute to maintaining the volume and quality of scientific output (OECD, 1998).

Electronic publishing

While informal means of communicating among scientists have become more salient with the use of ICT, these technologies have also transformed scientific publishing, the formal means of diffusing research outputs (*Bourguignon; Leeflang and Gilmore*).

Before the advent of electronic publishing, the researcher who wanted his work to be made broadly available to colleagues had to publish it in a paper-based periodical, just like any other author. In order for his article to be published, the author assigned copyright to a publisher who would protect the author from the theft of authorship and content of his article and who would, with the copyright, recover the cost of his substantial investment and make a profit on it. Access to articles by potential users was via research libraries that subscribed to the periodicals. As the prices of paper periodicals rose, the number of libraries that could afford to buy them diminished, increasingly limiting access. Consequently, the work could not be disseminated as broadly as scientists would have wished (Harnad and Hemus, 1997).

Potential impacts of ICT on scientific publishing[2]

The opportunities offered by ICT have led scientific publishing into a period of revolution, affecting all aspects of publication. Some of the characteristics of this revolutionary process are:

– *New document production tools:* authors themselves can produce high-quality manuscripts using sophisticated word-processing software such as TeX.

- *Enlarged access:* stored documents are accessible from anywhere in the world, radically altering the notion of distance.
- *Different economical constraints:* the selection process through which manuscripts were accepted by journals until now was directly related to production costs; this is no longer valid with the advent of electronic publishing.
- *Storage:* storage is no longer a matter of volume, it has become more sophisticated, *e.g.* appropriate tools must exist to allow files to be transferred from one operating system to the next generation.
- *Need for new tools: e.g.* appropriate search engines and, with the availability of primary literature on line, software which allows users to switch from one language to another.
- *Types of document:* animated images will soon be storable.

The impact of the new technologies on a given science depends on the weighting of the above factors.

The main vehicle for electronic publishing is the Internet, a network rooted in a philosophy of free access. This raises the question of control over the documents made available on the Internet (copyright, authentication). In addition, the new technologies are changing the professional skills required of those involved in the distribution of information; they now need to become experts in (tele)communication.

Some emerging trends can be highlighted:

- Greater speed of circulation of pre-publications, partly due to the fact that scientists themselves can now prepare their manuscripts to high standards of quality.
- Attempts to launch cheaper forms of publication, mainly in electronic form.
- New possibilities for rapidly, and almost effortlessly, gathering a large number of references, thanks to the electronic format of bibliographical databases.
- Less reluctance to publish in specialised reviews since online publications can be more easily located.

Major uncertainties

The future of electronic journal publishing is rather uncertain. One of the uncertainties relates to the time and cost savings that can be realised with electronic publishing. The timesaving of e-publishing is not as advantageous compared to print as is sometimes thought. While lead times between the submission and publication of a text may be significantly shortened with electronic publishing, the end result depends on the character of the electronic publication and on the scientific discipline (Kling and MacKim, 1997).

Estimates of the cost savings generated by e-publishing also differ significantly. The additional cost of providing online access to existing paper publications are generally agreed to be in the 20-30% range, including the cost of providing access to and maintaining a digital archive but excluding costs to the user (ICSU, 1998). Nevertheless, some studies suggest that savings are heavily dependent on the number of subscribers. In the case of purely electronic publications, there are indications that these may be produced for about the same cost as a printed version but experience in this area is limited (ICSU, 1998). According to Stevan Harnad, long-time editor of refereed paper and electronic journals and recent creator of an electronic archive on cognitive sciences, savings by publishers who have launched an electronic-only publication may be as high as 75% compared to the current cost of paper publication (Harnad and Hemus, 1997).

In spite of the numerous achieved and potential benefits, electronic publishing of primary publications and secondary databases has until recently constituted an up-to-date, rapidly accessible complement to print rather than a complete substitute. Revenue from electronic products still constitutes less than 15% of the total revenue of traditional commercial publishers of science, technology and medicine (*Leeflang and Gilmore*).

Applications which are successful in certain disciplines may not necessarily be realisable elsewhere, since collaborative practices, uses of pre-journal publication formats and uses of research journals differ across disciplines (Kling and McKim, 1997). One of the innovative applications of ICT to research – the development of the Physics E-Print Archive by Paul Ginsparg – is frequently cited as a model for global electronic pre-print archives throughout science. However, high-energy physics is characterised by particular research practices. Research evolves around a limited number of expensive instruments and involves large collaborative research groups working together over long time horizons. The research has typically been extensively reviewed by the time it is submitted to a journal, the time to publication is short and pre-prints can be made available simultaneously with little risk of plagiarism.

This is hardly the case in biological research, which is quite fragmented and involves many small research groups and individual researchers. Research facilities are common and relatively cheap. Biological research is also easier to extend and copy. Researchers in this field are therefore reluctant to share research prior to publication. In areas of biology, such as cancer or AIDS research, that are closely linked to commercial applications researchers often work with the private sector. The fact that such work can be lucrative and is highly competitive explains why researchers are unwilling to share research methods, materials and results. Publication in biology is therefore centred on journals, and pre-prints are quite rare.

Broad diffusion of and access to research outputs highlight questions concerning control of quality (peer review) and control of documents made available over the Internet (validating authorship and date of publication, copyright). An awareness of the (lack of) quality of the increasing amounts of data and information to which ICT provide broad access is an absolute necessity for users and for the assessment of individual researchers and of the science system. While both publishers and scholars agree that the need for peer review persists as it constitutes a quality label for research output, e-publishing has drawn attention to the need for new methods to implement peer review (which is typically part of the publishing process in printed refereed journals). The International Council of Scientific Unions (ICSU) has recommended that scientific societies and publishers establish guidelines to guarantee the quality and accuracy of the refereeing process in a digital environment (ICSU, 1998).

Copyright issues and intellectual property rights of authors are in conflict with the demands of some commercial publishers. There continues to be a significant gap between the views of publishers and producers, on the one hand, and scholars and librarians, on the other, as to what constitutes "fair use" of copyrighted documents, images, film clips on campus networks and in online courses (*Chronicle of Higher Education*, 1997). Copyright owners are afraid of losing control once their material is on the network, whereas academics want to maintain their right to use material freely for educational and scholarly pursuits. Differences in points of view have led to suggestions that researchers should retain copyright to their work and give journals a limited licence to publish papers without owning them so as to be able to disseminate their own findings through the World Wide Web (*Chronicle of Higher Education*, 1998).

Publishers want to protect their investments when fulfilling their role in enhancing the flow of information, including data, to the scientific community worldwide and to set boundaries for fair use. Generally assembled by scientists in the course of their research, factual databases frequently end up in the public domain. Those that continue to exist beyond a few years are usually subsidised by national or international institutions. Privatisation is unlikely given the extent of subsidies in this area in various parts of the world and the price tag it would entail for end-users. Publishers call for interactions between public and private actors, notably in relation to funding, to meet the new challenges raised by ICT including the conditions under which subsidised and unsubsidised projects on factual databases compete in international markets (*Leeflang and Gilmore*).

Electronic publishing in science is being held back to a certain extent by the reticence of researchers. The reputation of the Internet as a medium containing material of dubious quality has done nothing to persuade scientists that it is a suitable medium for disseminating serious work. As for the distribution of their work electronically before publication, they are afraid that publishers may consider this as prior publication and refuse to publish it. Furthermore, recognition and advancement is still very much based on publishing in established journals, practically all of which are paper-based (Harnad and Hemus, 1997). Governments could alleviate these fears by extending research support grants to cover the costs of publishing findings in refereed electronic journals and by taking refereed electronic publications into consideration when assessing research publication productivity of universities to determine levels of support (Harnad and Hemus, 1997).

Current measures to ensure the long-term availability and archiving of digital research papers and publications are inadequate. It is essential that publishers and scientists collaborate to establish basic principles covering long-term availability and archiving. In addition, government funding is essential for creating and maintaining an archive of electronic publications in science (ICSU, 1998). In disciplines where the broad distribution of digital pre-prints is feasible, governments may wish to accelerate the transition of refereed research journal literature to electronic media by subsidising the development of electronic pre-print archives.

Digital libraries

Rapid advances in ICT have made it possible to handle digital data and information in large volumes at ever-increasing speeds and falling costs. Storing, filtering, processing, compressing and retrieving digital data for analysis and retransmission have been tremendously facilitated. Digital databases constitute increasingly powerful ICT tools, available to a growing base of scientists and engineers, that enable new ways of working by increasing researchers' ability to access information and significantly improve the efficiency of information-based work. Internet tools, in particular, have made information more readily available as information service providers have started moving to Web-based systems. Methods of pricing information are also evolving and their impacts need to be considered as they may limit access to primary sources of information.

Digital libraries store and manipulate large collections of material in increasingly electronic form and allow rapid access to these resources when and where users desire it. Their development is closely linked to that of network information systems and the ability to integrate net-based services with traditional services. They are increasingly seen as projects where established libraries co-operate to provide networked access to collections of disparate resources on disparate systems, for example both paper-based and digital, through a single interface for the greater benefit of users and as a way to leverage investments by sharing resources. This evolution is leading to a need for training and organisational changes that are essential to maximise the benefits of ICT investments.

Researchers have repeatedly stressed the need to maintain the central role of libraries, when turning towards digital technologies, as the principal tools that enable them to advance from past knowledge into the future. More thinking and efforts are required to efficiently achieve this link in an increasingly digital research information context. This is especially true where all past knowledge cannot be digitised. Even retrieving full-text articles from electronic research libraries can be quite laborious at present. Also, appropriate means of storing and accessing the enormous amounts of data collected are severely lacking in some fields of research, notably in the social sciences.

A number of governments have launched projects to develop digital libraries based on varying systems of access and digitisation policies. Denmark (The Danish Electronic Research Library), France (*Bibliothèque Nationale de France*), the United Kingdom (Initiatives for Access of the British Library) and the United States (the US Library of Congress) constitute a few examples. The article by *Thorhauge* describes the co-operative project of three ministries in Denmark (Research, Culture and Education) to provide researchers with easy access to the entire content of networked research libraries in Denmark from any computer. A five-year grant of US$32 million provides funding for the project. The project covers the national infrastructure, the library infrastructure, the digital resources and the user facilities.

Digital library research and development projects aim at making the collections of published resources in individual traditional research libraries and increasingly large volumes of electronic information diffused on line, available to researchers on their computers. Projects to develop digital libraries and to carry out research on digital library issues are rooted in the information infrastructure initiatives of the early 1990s. Amongst these, the Digital Library Initiative of the United States, for example, first focused on the infrastructure issues raised by large-scale digital libraries. The National Science Foundation, Advanced Research Projects Agency and National Aeronautics Space Administration provided roughly US$1 million of funding per year for each of the six technology development projects during the first four-year phase of this initiative.

For the second phase, emphasis has shifted from technology to content and users. The need to promote university-industry research partnerships, particularly with the online technology industry, has also been recognised. The purpose is not only to benefit from the additional funds that companies may be willing to contribute and to speed up commercialisation but, more importantly, to avoid being taken over by unexpected rapid commercial developments.

Among the components of digital libraries, content seems to pose particular problems. Although the potential of ICT is to permit access to all information on a certain subject to users in any location, this is not yet the case. At present, electronic versions of traditional publications and new kinds of retrieval are being made available. Methods and funding for the conversion of full-text articles require further consideration and, particularly, international co-operation to avoid that the same tasks be carried out in parallel in many countries. The range of materials currently considered as constituents of digital libraries does not include databases that contain raw data and it is not clear whether it should and whether that is a concern for governments or for library professionals. The extent to which libraries should be responsible for their own adaptation and conversion into the digital world using their own funds or with additional public grants also raises questions. So does funding for research, development and implementation of tools for indexing and searching. Building digital libraries is a complex and expensive task that necessitates a long-term commitment.

Databases

Bibliographic (secondary) databases providing information on published research in the form of citations, abstracts and various indexes to scientific research literature allow scientists and researchers to identify published articles appropriate to their needs. They range in size from more than 20 million records in the On-Line Computer Library Centre to millions of entries in online databases for specific disciplines. ICT has expanded delivery methods for these databases and has created better options for storage, search and retrieval, as can be seen in the case of mathematical databases. Scientists in many fields now produce factual databases that are available via the Internet to scientists around the globe. The Internet also provides new opportunities for scientists in different countries to combine local data sets into global ones, particularly in research projects that require data from around the world, as in biological and Earth-related sciences (Human Genome Project, International Geosphere-Biosphere Programme). Large-scale factual databases could also transform social science research, but are currently lacking in those disciplines.

Mathematics: bibliographical databases[3]

Mathematics is particularly demanding of bibliographical information because written documents play a very specific role in this field. A rigorous proof – the ultimate goal in this discipline – is equally relevant today as it was in the past. Published mathematical work contains all the information necessary for the competent reader to thoroughly check the data *(Bourguignon/EMS)*. This is obviously not the case for experimental work. New concepts are also developed in mathematics, however they do not automatically render irrelevant the concepts used in the past (which is often the case in other fields). Although the need to refer to previously published literature is not as acute in all disciplines, the case of mathematics highlights many issues and principles which should be considered when taking advantage of the technical opportunities offered by ICT in the development of disciplinary databases.

Mathematicians make extensive use of two bibliographical databases as research tools, *Mathematical Reviews* produced in North America and *Zentralblatt-MATH* in Europe, bearing on the entire mathematical literature from the 1930s to the present day. These are increasingly accessed through an international system of mirror servers on the Internet (Berlin, Strasbourg, New York). Local servers, either regional or serving specific institutions, offer high-performance access (via CD-Rom). Each notice includes an analysis written by a mathematician specialised in the subject. A network of over 5 000 scientists contributes to this body of reviewers.

The possibility for researchers to easily access these databases to check a statement, evaluate the production of a researcher or get an idea of scientific links between two subjects is often at the origin of a major discovery. Running the two databases costs about US$1 million per year. In contrast to the 60 000 notices added to each database every year, the digitisation of an earlier archive covering 200 000 notices from 1868 to 1942 is being financed from public funds since commercial involvement would have resulted in a product that would have been far too expensive for widespread use.

There are several criteria for evaluating the quality of a database. Its comprehensiveness within the boundaries of the field is very important. The quality of its reviewers is essential – reviewers are usually unpaid, although they may receive some fringe benefits from the publishing houses involved – and inclusion of experts from around the world is an advantage. The professional competence of the team of editors distributing the articles to the reviewers is also significant, as is the speed with which articles are handled (the standard delay in mathematics is between six and eight months after publication in a journal). In addition, the classification and organisation of material in a database determines to a large extent its usability both for searches, which are one of the principal advantages of new electronic formats, and for consultations.

Retrieval software is critical for the efficient use of the database; it must satisfy a number of requirements, including user friendliness, efficiency, flexibility and robustness. The database needs to be fast but should also be unambiguous as to how authors' names, titles of journals, etc., are entered into the base. This can only be achieved through the use of a specific thesaurus for these categories. It should nevertheless be able to recognise first names that are entered in different ways or transliteration from other alphabets or ideogrammes that has varied over the years. It should be able to interact with existing standard software and should not be overly syntax-sensitive.

Another key question concerns the control of databases. Since the content of this type of database comes from the scientific community itself, it may be appropriate that learned societies be involved and hold the copyright, as in the case of maths. There could be a danger, however, that one group might try to impose its scientific views on a particular theme. Databases represent a powerful lever and this is a good argument for a competitive system, unless appropriate measures are taken to preserve the independence of the database management from lobbies.

Concerning the provision of access to developing countries, a licensing arrangement adopted for mathematical databases is worth noting. An institution in an better-endowed country can be "twinned" with one in a developing country, which only pays a small percentage of the normal price. The two institutions share part of the services such as training. Although the additional income to the database owner tends to be marginal, by making the products affordable the owner can reach new users.

A critical step to further improve the services offered by bibliographical databases relates to the interface with primary literature. Establishing hyperlinks between review databases and the full–text articles will tremendously increase the efficiency of reference searching. However access to full-text articles remains difficult at this stage.

Social sciences: large-scale databases

A significant number of science-related issues that governments of OECD Member countries have to address today relate to the social sciences, the spectrum of disciplines which broadly deal with the behaviour and interactions of people and social institutions. Sustainable development, changes in the labour market, family instability and population ageing are but a few examples of major problems faced by Member countries to the resolution of which the social sciences can contribute.

Yet, the social sciences have generally lagged behind natural sciences in terms of budgetary appropriations – particularly in times of crisis, in terms of policy impact and in terms of attractiveness as a field of study. To a certain extent, the social sciences and humanities receive less funding than other scientific disciplines because they do not require the same expensive capital investments (instruments and facilities). Nevertheless, the social sciences have their own research infrastructure requirements, of which large-scale databases are one. Dealing with issues that relate to individuals, their families, their workplaces and their organisations makes it necessary to survey large numbers of people on many different dimensions and to store survey data in large-scale databases (OECD, 1999).

Inadequate data and insufficient possibilities for data sharing have been identified as major obstacles to integrating the social sciences into policy making. In most OECD countries, multiple and fragmented databases containing a wealth of information relevant to social science research and analysis have to date been accessible only to relatively narrow circles of users. In particular, these databases have remained inaccessible for many types of secondary analysis that are required for developing policy options for consideration by governments. Most of these data sets originated from different, relatively small and costly national surveys which produced data that are not necessarily compatible with each other *(Bainbridge)*.

The new ICT provide the technical feasibility to build comprehensive electronic databases in many scientific disciplines. ICT tools such as the Internet can make the scattered social science databases accessible to virtually all qualified researchers. They may also constitute a less costly means to carry out new surveys. In addition, ICT can serve to apply experimental methodologies to the study of large-scale networks of human interaction to reach and analyse responses from a far greater number of subjects than was possible before the development and extensive use of these technologies. Studies of markets, election systems and social networks are cases in point.

The utility of social science databases will be strongly improved if they are international in scope; that is, if analogous databases in different countries are linked and if researchers in different countries can access them. Government agencies in certain countries have recently begun to expand their survey support to cover the dissemination of results via the Web. Researchers and students from different parts of the world can thus gain access to data that they can either analyse on line or download to their computers, as well as to published results, analysis tools and relevant references.

There are many challenges to the development of large-scale electronic databases in the social sciences. Emphasis must be put on the quality of data, including reliability and standardisation, in addition to data storage and retrieval. A challenge that is specific to the social sciences (and medical sciences), especially when maximising access by social scientists and government policy makers to personal data, is the protection of privacy. Similarly, issues of confidentiality may arise in surveys targeting business enterprises. Questions on intellectual property rights in relation to databases also need to be addressed. Last, but not least, it may be necessary to nurture a different set of cultural values for collecting and sharing data (OECD, 1999).

Addressing the technical and policy issues involved in expanding and improving the efficiency of electronically linked, widely accessible, international databases containing data relevant to social science research in a way that will lead to the implementation of appropriate policies and procedures requires consultations among governments. To that end, a Workshop on Large Scale Infrastructure for the Social Sciences, to be hosted by Canada in the fall of 1999, will examine the relationship between infrastructure development in the social sciences and its potential impact on research practices and productivity.

General policy orientations

At the Sintra Conference, governments[4] emphasised two broad opportunities presented by the use of ICT as a tool for science. One is to foster collaboration, not only within the science system but also between the science system and the economy and society as a whole. Better links between the science system and business, including partnerships between universities and business enterprises for ICT development and use, could enhance the innovation process. Stronger links between the science system and society would reduce the relative isolation of the science system, thought to be one of the reasons behind limited public understanding and support of its activities. This link is also intended to achieve a richer science and technology culture, using ICT access by households to improve opportunities for lifelong learning. Nevertheless, governments recognise that high priority must still be given to the specific needs of the research community, which is often the driving force in the development of ICT and related applications in business and society.

The other broad opportunity is the use of ICT as a tool to better integrate the scientific communities of less developed countries into the Global Research Village, ensuring that the latter does not exclude developing countries. The principal means for researchers in these countries (who have little opportunity to influence their decision makers) to engage in international scientific co-operation is direct interaction with scientific communities of developed countries through the Internet. Empowering researchers in developing countries with the requisite information technology infrastructure and human resource base to participate in international research projects is therefore crucial.

Although considerable opportunities are offered by ICT to the science enterprise, a number of issues – some of which had already been identified earlier but on which progress has been slow – require action by the public authorities. There was general agreement at the Conference on the broad tasks that governments will have to fulfil in order to maximise the benefits of ICT for science:[5]

– Ensure that a high-performance information and communication infrastructure is available to scientists at all times. To this end, governments must provide funding to develop high-speed research networks especially when commercial networks do not provide the required capacity at a reasonable price. They must also provide leadership and funding for research and development in components of the infrastructure for which researchers constitute the principal source of demand and which the business sector is consequently unlikely to develop on its own.

– Promote the training of all scientists in the use of ICT so that they are able and willing to make the best use of ICT tools. Although scientists are among the most skilled workers in the OECD area and university education gives increasing attention to developing ICT skills, scientists may not be sufficiently aware of the potential of ICT or necessarily know how to use ICT in the most efficient way. Collaborative work, in particular, requires a substantial investment in learning to use ICT. Efficiency of use also increasingly depends on the availability of technical support. In addition, information professionals must be duly trained to carry out important support work for researchers, including the organisation of, manipulation of and rapid access to relevant information through new network-based services.

– Stimulate the provision of infrastructure-based services and applications that will maximise the benefits of ICT for science and make it possible to improve the links between science, business and society by implementing appropriate regulatory frameworks conducive to rapid and efficient developments. Governments of many OECD countries have already undertaken regulatory reform affecting the communications industry, but more remains to be done.

– Ensure that realising the benefits of ICT for science is not hampered by legal frameworks governing international collaboration in research and control over online information, including intellectual property rights, transborder flows of data, privacy protection, information security, and authentication. Governments must continue to pursue greater compatibility between national systems. The further development of electronic publishing in science, digital research libraries and databases depends on due consideration of the interests of the science community in the current debate concerning the respective rights of authors, publishers, librarians and users. In addition, governments, in co-operation with the scientific establishment and commercial entities, can contribute to the resolution of uncertainties related to access to infrastructures and databases, usage costs, quality control, recognition and long-term availability and archiving of research outputs diffused on line.

– Make a more concerted international effort to reduce the significant gap in the present level of infrastructure development between developed and developing countries to realise a global infrastructure and the Global Research Village. Disparities in the physical infrastructure that constitutes the basis of all ICT applications in science are striking when the developing world is considered. A few projects, such as the global Oxygen Network and APAN in the Asia Pacific Area, are good examples of what can be accomplished in this area, but much remains to be done and a lot more time may pass before disparities are reduced.

– Strengthen international co-operation in general to avoid duplication of efforts and funding in areas such as the conversion of research outputs diffused on paper into digital media for inclusion in digital research libraries.

– Develop internationally comparable indicators for a quantitative assessment of ongoing developments and performance, including return on investments in this rapidly evolving area. Diversity in the way each discipline, country and region views and is able to organise infrastructure, content and access seems to be the dominant characteristic in the evolution of ICT use in science, and needs to be measured.

There are many other potentially beneficial developments for science that may at best be significantly delayed without additional funding from governments. These generally relate to the development and long-term availability of digital content. Digitising paper-based material for the development of digital research libraries, archiving electronic publications, improving the life-span of important/unique factual databases in certain disciplines as well as supporting the creation of large-scale factual databases in the social sciences constitute some of the calls for public funding.

Notwithstanding the needs for further action recognised by governments, maximising the benefits of ICT on the science system is difficult. The technologies are continuing to evolve at a very fast rate. Their potential and limitations are still uncertain. Policy action towards the science system must be taken in full consideration of these uncertainties, and must remain flexible.

Viviane Bayar
Jean-Éric Aubert

NOTES

1. The information on this region is based on the presentation entitled “Internet and Science in the Asia-Pacific Area: A Japanese Viewpoint”, made by Mr. H. Ishida of ASCII Future Labs, Tama Art University, at the Sintra Conference.
2. This section is an extract from the presentation entitled “Maximising the Benefits of Information Technology for Science: The Case of Mathematical Databases”, made by Mr. J.-P. Bourguignon, President of the European Mathematical Society, at the Sintra Conference.
3. This section draws heavily on the presentation entitled “Maximising the Benefits of Information Technology for Science: The Case of Mathematical Databases”, made by Mr. J.-P. Bourguignon, President of the European Mathematical Society, at the Sintra Conference.
4. Mr. José Mariano Gago, Minister of Science and Technology of Portugal, who chaired the Conference, and Mr. Jon Trøjborg, Minister of Research and Information Technology of Denmark.
5. As expressed in the conclusions adopted at the end of the Sintra Conference.

BIBLIOGRAPHY

CHRONICLE OF HIGHER EDUCATION (1998),
"Scholars Seek New Copyright Rule to Ease Dissemination of Research on the Web", 11 September.

HARNAD, S. and M. HEMUS (1997),
"All Or None: No Stable Hybrid or Half-way Solutions for Launching the Learned Periodical Literature into the Post Gutenberg Galaxy", ftp://ftp.princeton.edu/pub/harnad/HTML/harnad97.hybrid.pub.html, 16 December.

ICSU (1998),
"Executive Summary Report, ICSU Press Workshop on the Economics, Real Costs and Benefits of Electronic Publishing in Science – A Technical Study", published as supplement to *Science International Newsletter,* No. 68, Paris.

ICSU (1998*a*),
"Meeting Reports", AAAS/ICSU Press/UNESCO Workshop on Electronic Publishing in Science, No. 69, Paris.

JAPAN SCIENCE AND TECHNOLOGY AGENCY (1997),
White Paper on Science and Technology 1997, Tokyo.

KLING, R. and G. McKIM (1997),
"Electronic Media and Scientific Communication: Opportunities, Complexities and Limitations", consultancy paper for OECD, Indiana University School of Library and Information Science, September, mimeo.

OECD (1998),
"The Global Research Village", http://www.oecd.org/dsti/sti/s_t/scs/prod/online.htm.

OECD (1998),
Science, Technology and Industry Outlook 1998, Paris.

OECD (1999),
The Social Sciences at a Turning Point?, Paris.

OECD (1999),
Communications Outlook 1999, Paris.

A GLOBAL RESEARCH INFRASTRUCTURE FOR THE GLOBAL RESEARCH VILLAGE: THE EUROPEAN PERSPECTIVE

Table of contents

This article was written by Hans-Peter Axmann and Sabine Payr. Mr. Axmann is head of the Computer Science and Networking Division of the Federal Ministry of Science and Transport in Austria. He is also a lecturer of computer science at the University of Vienna. He is currently chairman of ENPG, the European Group for Policy Co-ordination for Academic and Industrial Research Networking (e-mail: hans-peter.axmann@bmwf.gv.at). Ms. Payr is a researcher and consultant in telematics and multimedia applications in higher education. She has been a university lecturer in computer graphics and multimedia design, and works as research scientist for the Research Association for Computer Science. She served as secretary of ENPG until December 1998 (e-mail: s.payr@fgi.at).

OECD 1999

I. INTRODUCTION

Industrialised countries are heavily reliant on innovative and successful research and have a vested interest in seeing research produce a wealth of excellent results. In most cases, these results lead directly or indirectly to new industrial products and thus contribute to economic development. Research, then, is also an important precondition for the creation of new, and the preservation of existing, employment, for the maintenance and improvement of the quality of life in the countries of the European Union, and the answer to the global challenges raised by technology and economic competition.

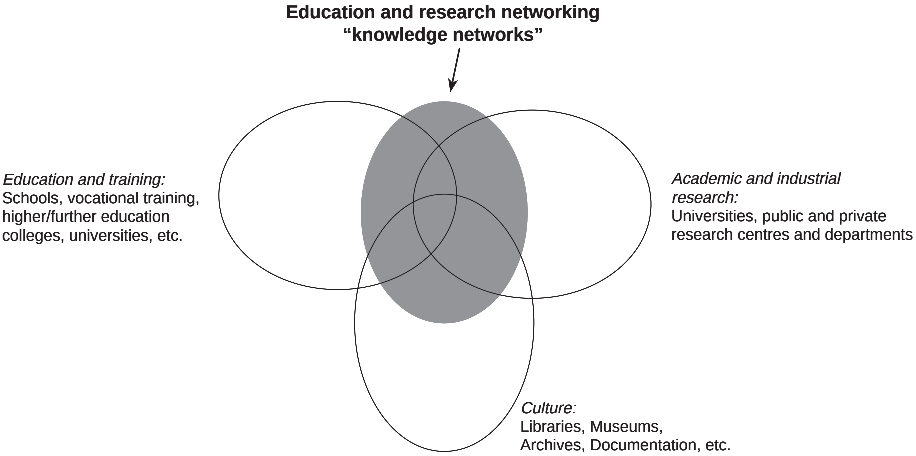
Governments of industrialised countries are therefore responsible for the creation of an environment that is conducive to research and innovation. Only in a research-friendly atmosphere can research flourish and be successful.

Different approaches and actions can be taken to achieve this research-friendly environment. These range from tax incentives for research activities to indirect, and even direct, government involvement in issues pertaining to research (*e.g.* the pursuit of national policies for the advancement of research and technological development; the stimulation of research activities in industry, including both large enterprises and SMEs; the setting up of mission-oriented research institutes).

High-performance networks for electronic communication have become indispensable to this environment, and almost all research activity depends heavily on them. Today, research networks are an important tool, not only in all academic disciplines – ranging from natural sciences to medicine and the humanities – but also in the networking of libraries, museums, educational and cultural institutions. Taking into account the “mission” aspects of these institutions, it would seem to be more appropriate to speak of *knowledge networks* (Figure 1). Since this terminology is not yet widely accepted, in this article we use the term “research networks”.

Public funding has enabled the rapid development of a number of research networks at the national level. However, lack of an appropriate designated authority has meant that there has been no equivalent development at the European level. This has resulted in a heterogeneous European network structure, to the disadvantage of those countries lacking an adequate national network infrastructure. Their scientific development is already adversely affected by insufficient communication facilities, so there is a risk that the gap between these countries and those that are better equipped will become wider as time goes on.

Figure 1. **Knowledge networks**



Source: Authors.

The promotion of research networking at the European level has been limited, so far, to the various Community programmes for research and technological development. Indeed, the first truly pan-European research network IXI was established in 1990 within the framework of the Eureka project COSINE, which ran from 1987 to 1993. This network developed into EuropaNET in 1992, a 2-Mbps multi-protocol network supporting both X.25 and IP traffic. EuropaNET, in turn, was replaced by TEN-34 in 1997, a network nominally running at 34 Mbps, which could only be established with the financial support of the EC's Fourth Framework Programme (4FP).

As the importance of networking for research and technological development became more widely recognised, the issue was taken up in the 4FP. Actions in favour of research networks have become possible in the form of research projects. However, two fundamentally different aspects of research networking have to be distinguished:

– *Research in networking:* development and demonstration of advanced network technologies, services and applications.

– *Networks for research:* application of advanced networking technologies in all disciplines and fields of research.

The Fourth Framework Programme did not take these two aspects into account, neither did it clearly distinguish between them. Thus, supporting networks for research (the second aspect) was only possible indirectly, via research in networking (the first aspect). This is one of the reasons why the review panel of the Telematics Applications Programme characterised the research sector as "peculiar" and concluded that "the infrastructure work is of such importance that it could stand alone as a special activity in its own right." (Telematics Applications Programme, 1996).

II. THE FIFTH FRAMEWORK PROGRAMME

The European Commission's proposal for the Fifth Framework Programme (5FP) (European Commission, 1997) indicates that these and similar comments have now been taken into consideration. All the thematic programmes proposed so far contain a specific action line for research infrastructure. The programme dealing with the Information Society, in particular, lists high-bandwidth networks as a necessary component of the research infrastructure.

In the process of defining 5FP, it has become apparent that a differentiated approach has to be adopted in relation to the way in which the listed actions will be implemented and which instruments will be used to pursue the stated goals. Following the distinction made above, it is clear that:

– Research in networking has to remain a priority topic. It can be fairly well accommodated within the regulations of the framework programmes, through the usual instrument of project-based support. However, increasing globalisation requires that more attention be given to the worldwide acceptability of results and products in order to avoid isolated solutions for Europe.

– Networks for research need to be considered as an enabling technology as well as a (pre-) necessary infrastructure. As an essential tool for the promotion of co-operation among member states and for research and technological development as a whole, research networks should be seen as an activity which crosses all programmes and thematic lines, since it is quite obvious that research networking also serves all other themes and actions in the 5FP.

In order to proceed in the best possible way, a further distinction is necessary. The long experience of the research and education communities served by a European backbone network suggests that two different requirements need to be addressed. The vast majority of end-users clearly require a stable and efficient "production" network for their everyday work. On the other hand, a relatively small group of users, particularly those involved in research in networking and the development of new networked applications, but also those running trials and making innovative use of existing applications, needs an infrastructure for experimenting and testing – a “test-bed” network.

It has been argued in the past that these different patterns of use warrant the establishment of separate networks. On closer examination, however, it becomes clear that production and test-bed networks have so many features in common that it makes good sense to set up both on the same physical infrastructure: both networks need to be managed, therefore, the management should be entrusted to those who have proved capable of managing networks for scientific use, *i.e.* the NRENs and Dante. Although Internet traffic can presently only be handled on a “best-effort” basis, even “normal” users already require guaranteed bandwidth and quality of service for certain applications. The management issues associated with this are similar to those of setting up dedicated connections for individual test applications. Furthermore, an application which has been successfully tested in a “secure” environment can be more easily transferred to a “real life” environment if both environments use the same underlying physical infrastructure. Finally, this solution would be financially far more economical.

The instruments for project-based support, because of their limited timeframe, are, however, not adequate to reflect the needs of research networking. A more direct and flexible, taskforce-like approach should be adopted, ensuring the necessary support for the duration of the 5FP. Furthermore, there now seems to be a consensus that support for infrastructure has to be handled as a procurement activity.

A fundamental pre-condition for this is, however, that the necessary funding for *a pervasive high-performance pan-European network* can be secured for at least the time period covered by a framework programme. Since this is in the interest of the national authorities as well as the European Commission, a co-funding model has been suggested (Axmann *et al.*, 1997). Following the example of, and taking into account the lessons learned from TEN-34, a shared-funding model would seem to be most appropriate. Under this scenario, EU member states would contribute 60% of total costs through their National Research and Education Networks (NRENs), while the European Commission would contribute the remaining 40%. The same model should also apply to all the European research institutions to enable them to participate in this infrastructure.

Based on input from Dante, it has been estimated that a pervasive network with a transmission speed of 155 Mbps between all member states (with two lines each for resilience) would cost some ECU 125 million per year. Using the 40%:60% cost-sharing model described above, the EC contribution would amount to ECU 50 million per year. Over the four years of the 5FP, this would amount to ECU 200 million. Compared with the original EC budget proposal of approximately ECU 4 000 million for the whole Information Society (IST) Programme, the contribution for the network infrastructure would be 5%, which would seem to be a very acceptable overhead. However, budgetary restraints in most member states have reduced this figure significantly in the European Council's common position as accepted in February/March 1998. Furthermore, the percentages used in the indicative budget breakdown of 8 June 1998 show that the Council is prepared to allocate 4% of the IST budget for infrastructure support, while the EC has so far only come up with 3%. In real terms, the equivalent amounts are ECU 135 million and ECU 118 million, respectively. This is clearly insufficient to achieve the desired goal.

The recent conciliation process between the Council and the European Parliament has resulted in the allocation of ECU 3 600 million for the IST Programme. Application of the above percentages of 3% (EC) would give ECU 108 million, and 4% (UK presidency suggestion) would give ECU 144 million; it is clear that the situation has not dramatically improved.

Finally, the Council of Research Ministers, when formally deciding on the 5FP, agreed on 22 December 1998, under the Austrian Presidency, to allocate 4.5% of the budget of the IST Programme to support for research networking; this amounts to ECU 161 million. EC officials have expressed their confidence that additional money can be found elsewhere. This looks promising, although it is still unclear whether the whole sum will be allocated to the network infrastructure and how much of it will go to other activities.

III. THE PARTIES INVOLVED

Past experience has shown that it is the research and higher education community that needs the most advanced network infrastructure. However, even a year after the liberalisation of the European telecommunications market, this infrastructure is not readily available as a market commodity. Therefore the usual market mechanisms do not apply. The need for a continuing, public (co-)financing of research networking has been recognised by the member states as well as by the European Community. The present 5FP will at long last provide the opportunity to finance infrastructure. The Commission will be responsible for the co-financing and the mechanisms to implement it.

It is crucial that all parties involved in European networking in higher education and research co-ordinate and co-operate in the planning and implementation of this infrastructure. At present, the following actors on the scene of research networking can be distinguished:

- The national governments or their respective agencies: They are responsible for research networking on a national basis, as outlined above. There is clearly a need for national governments to co-ordinate their activities at the European level, in particular where European and intercontinental connectivity is concerned.

- The European Group for Policy Co-ordination for Academic and Industrial Research Networking, ENPG, has been established by most European countries as their instrument of co-ordination of national policies and can thus be considered as the player representing national governments and agencies.[1]
- The Commission of the European Union, CEC, can act within the framework of the Treaty (or Treaties) to promote research and support research infrastructure at Community level.
- The National Research and Education Networks, NRENs, are the operators of the networks that serve research and higher education at the national level.
- TERENA is the organisation of European research network operators. Its main goal is to promote better international research networking on behalf of its members and their users. Therefore, its mission is to promote and participate in the development of a high-quality international information and telecommunications infrastructure for the benefit of research and education.
- DANTE is a not-for-profit company, set up in 1993 by European national research networks to organise international network services for European researchers.
- "Industry" has not, as yet, been a very active player in European research networking. Hopefully, the importance of action and the potential for industrial development and markets will be recognised as, indeed, some recent examples seem to indicate. Industry is not only needed as supplier of physical infrastructure but also has much to gain as partner. For it to become a serious partner, however, real or imagined conflicts between the needs of research networking and business interests will have to be reconciled. In the now largely deregulated European telecommunications market, industry is still in the process of reorganising itself. Once the new public network operators (PNOs) emerge, it can be assumed that these will have a vital interest in collaboration with the research institutions.

The four actors who are in the position to take the leading roles at present and whose co-operation is required are therefore: ENPG, the CEC, TERENA and DANTE. If each is prepared and committed to play its assigned role in co-operation with the other bodies, there should be no conflict between the parties. Their roles can be outlined as follows:

ENPG co-ordinates and brings together the national governments who, through their respective funding organisations, define their research networking policies. ENPG liases with the CEC to define a concrete action plan at the European level, including decisions on and commitments to financial contributions and the procedures for implementing the agreed actions. The role of TERENA is to bring the experience of the national networks and their commitment to the action plan. DANTE is to be mandated to carry out the work according to usual business procedures and/or special CEC rules and regulations as they apply. Thus, while the national governments through ENPG and the CEC are responsible for the harmonisation of strategies and the allocation of funds, TERENA, the NRENs, DANTE and its subcontractors will be responsible for ensuring that the work actually gets done.

IV. EXPERIENCES AND COMPARISONS

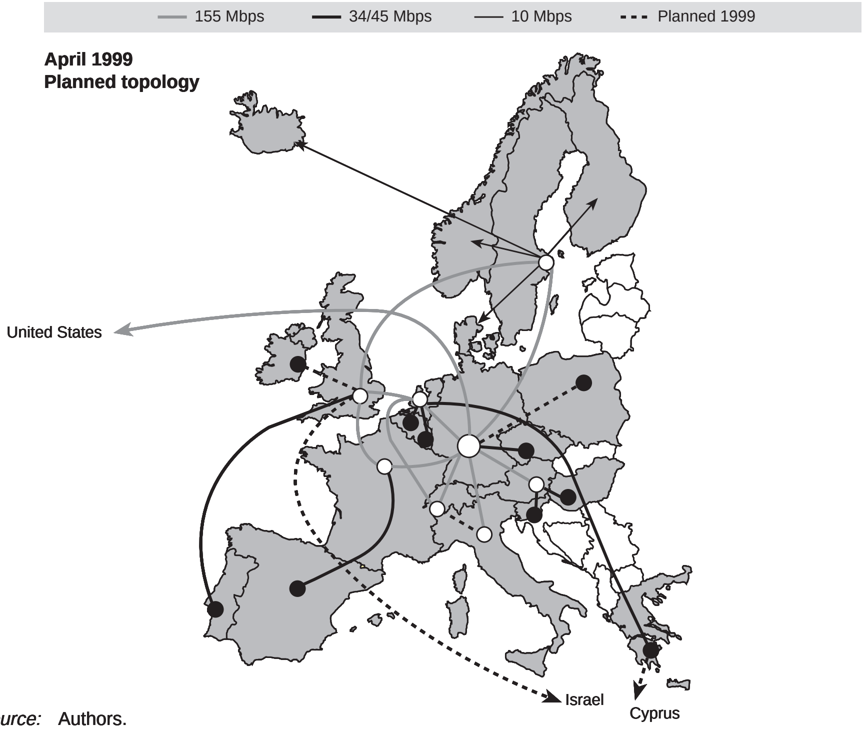
The situation in Europe

For the last ten years and more, Europeans have endeavoured to create digital electronic networks in order to enable researchers and educators to better communicate with each other, not only within the confines of a national state but also with their peers in other countries throughout Europe and the world. These activities have resulted in the establishment of national research networks in all the European countries and in the establishment of mechanisms to create and ensure internetwork services on a European scale. The way in which this was achieved was determined by the prevailing situation of monopolistic telecommunications operators in the European countries and the resulting difficulties of organising and funding the necessary European backbone infrastructure.

As mentioned above, the first European backbone was IXI, started in 1990, a purely X.25, OSI-oriented network with a transmission speed of 64 Kbps. IXI was replaced in 1992 by EuropaNET, which developed into a 2 Mbps IP-based network. In 1997, TEN-34 (nominally at 34 Mbps) was launched with the support from the European Commission's 4FP. Towards the end of 1999 it will be replaced by TEN-155, consisting of a core network running at 155 Mbps to which the NRENs can connect at various speeds as they deem necessary (Figure 2).[2]

On 1 January 1998, the telecommunications market experienced full deregulation in most European countries. As a consequence, it is hoped that competition will help to provide infrastructure and services which were not available up to now and will also bring down costs. These expectations are triggered by a market-oriented way of thinking. This line of thinking can be expanded, however, resulting in the notion that the provision and maintenance of infrastructure could be left completely to the PNOs since the market will provide what the customers need and the rules of the market will ensure competitive prices and fair deals for customers. However, current experience shows that this situation can, if at all, at best be expected some time in the future. As the telecommunications industry is only now

Figure 2. **Ten-155**



Source: Authors.

reorganising itself, business considerations take precedence over the needs of the research and education communities.

Due to the nature of the remit of the Framework Programmes, research was, for a long time, considered to be the sole legitimate reason for supporting activities in networking. Only recently has it been recognised that training and learning – and indeed all areas of education – are crucial to the future well-being of European citizens. That recognition has helped to change the thinking of policy makers, so that networking for educational purposes is no longer considered to be an "illegal", or at best a "tolerated" activity on a research network.

The situation in the United States

US Government involvement in networking was clear from the very beginning,[3] when the Department of Defense, through DARPA, triggered the development of the ARPANET which was operational from 1969 to 1990. In 1986, NSFnet, with a speed of 56 Kbps, was created. Two years later, this was upgraded to 1.54 Mbps and in 1992, it was again upgraded to 44 Mbps.

The same business-oriented line of thinking has been applied to the academic Internet in the United States, leading to the decommissioning of NSFnet in 1995. It is fair to say that, for many years beginning in 1987, the network services of NSFnet for research and education (R&E) were unequalled elsewhere. Its users have, however, noticed that the privatisation of that network and the frequent congestion of its commercial replacement have deprived many faculty members of the network capabilities required to support world-class research. It has been recognised that this unintended result has had a significant negative impact on the university research community in the United States.

To counteract this situation, leading universities reacted by creating the Internet2 initiative.[4] The primary goals of this project are to enable the development of new-generation applications, to re-create the leading-edge R&E network capability of the United States and to promote the transfer of results to the global production Internet. In order to achieve this, Internet2 members have set up a limited company, University Corporation for Advanced Internet Development (UCAID), whose mission is to provide leadership and direction for advanced networking development within the university community. At present, the vBNS backbone (very-high-speed backbone network service), which peers with other federally sponsored networks, is used by the Internet2 universities. vBNS was started in 1995 at 155 Mbps in the framework of a five-year co-operative agreement between the NSF and MCI; it currently operates at 622 Mbps.

In 1998, UCAID, together with Qwest, Cisco, Nortel and others, started the Abilene project.[5] This network, which became operational in January 1999 and started to deliver high-performance network services in February 1999, will offer UCAID members transmission speeds of 2.4 Gbps between core nodes, to be increased to 9.6 Gbps at a later date. This project and the Internet2 programme of development of advanced applications can be expected to have an enormous impact on networking.

Comparison and lessons to be learned

As mentioned above, the European countries have set up national organisations which are responsible for organising and/or operating their respective national research and education networks. These, in fact were really “university networks” from the start providing Internet services to their academic users. Over time these came to incorporate teachers, researchers and students alike. Although there are different policies in the various European countries as to who is eligible to use the national network (usually depending on who actually provides the funding for it), it can be said that a wide constituency of users is now served, very much in line with the argument made above about knowledge networks.

The US situation differs quite strikingly from this picture. There is no comparable national organisation responsible for networking and, perhaps as a consequence of this, not all institutions of higher education take part in the above-mentioned activities. Particularly since the decommissioning of NSFnet, most institutions have had to rely on market services. UCAID is set up to serve primarily (only) its 132 current members, *i.e.* less than 5% of all higher education institutions in the United States.

In view of the efforts to create a European Information Society, it is strongly recommended that availability of and access to a high-performance European backbone network should not be limited to a selected group of users. In this respect, European policy should aim to encompass a constituency which is as wide and as diverse as possible.

The commercialisation of the Internet has clearly shown that commercial providers follow a business model which does not (yet) take into account the special nature of infrastructure usage by the Internet and which is entirely based on common industrial and commercial principles, *i.e.* creating profits for their shareholders. The academic community's need for advanced applications to meet emerging requirements in research, teaching and learning trail far behind in the PNOs' list of priorities. The US experience of discontinuing support for NSFnet, clearly demonstrates that a similar development should be avoided in Europe. On the other hand, the European telecommunications operators should be encouraged to join forces with the networking actors mentioned above to accelerate the development of the Abilene project in Europe.

It has been said many times that a high-performance European backbone network is a prerequisite for the development of new and innovative applications. Many researchers have pointed out this chicken-and-egg situation: because there are no "killer" applications, there is no perceived need for high-performance networks; and because there are no high-performance networks, the new technologies which would require such networks are simply not being developed. The same scientists have argued that nobody would reasonably endeavour to develop novel applications if the appropriate infrastructure on which they could run is lacking: the appropriate network has to be put in place beforehand.

As demonstrated by the US example, a high-performance European backbone network is crucial if Europe intends to have a role to play in the development of the future Internet. The absence of such an infrastructure in Europe would be disastrous, with each NREN trying to forge its own agreement with UCAID to obtain access to Abilene and to take part in the overall development as "US-guest researchers" to avoid being left behind. Indeed, several NRENs are already in the process of negotiating with UCAID. The effect of this situation on European research and the further development of European co-operation in the Information Society cannot yet be gauged. There is, however, a danger that the idea of a European Research Space will be severely hampered and may never become reality.

The different approaches taken in Europe and the United States are summarised in Table 1.

Table 1. **Comparison of the European and US approaches to networking**

Networking: the European approach	Networking: the US approach
Infrastructure:	Infrastructure:
	1969 ARPANET commissioned by DoD
	1986 NSFNet created (56Kbps)
	1988 NSFNet at T1 (1.54 Mbps)
1990 IXI (X.25, OSI-oriented, 64 Kbps)	1991NSFNet at T3 (44 Mbps)
1992 EuropaNET launched (64 Kbps),	
developing into IP network at 2 Mbps	1995 vBNS at OC3 (155 Mbps)
1997 TEN-34 at E3 (34 Mbps) nominally;	currently at OC12 (622Mbps)
only one connection at 34 Mbps!	
1998/99 TEN-155 core at 155 Mbps	1999 Abilene at OC48 (2.4Gbps)
	upgrade planned to OC192(9.2Gbps)
Funding:	Funding:
– main funds from national governments	– originally government funded
– contribution from EC	– now paid for by users and sponsors
Purpose:	Purpose:
research (and education)	research AND education

From the above comparison, one can easily draw the conclusion that Europe – rather late – has entered into competition with the United States in trying to catch up on bandwidth. This raises the often-asked question: why does the education and research community need so much bandwidth? This is related to the second question: is there a “killer” application available which requires that amount of bandwidth?

There are two immediate answers to these questions. Firstly, the acceptance of networking services has been so high that network capacity would have to be increased by a factor of two every eight to nine months on average in order to meet demand. This creates considerable aggregate traffic load on backbone networks. Secondly, almost unnoticed by most, World Wide Web applications, which incorporate more and more graphic and animated elements as well as other innovative applications, simply need higher bandwidth in order to run satisfactorily. Table 2 attempts to calculate the bandwidth that will be needed to fulfil the expectations of the Information Society.

Table 2. **Why is so much bandwidth necessary?**

• Educational software (the delivery of interactive “learning” as opposed to mere “information”)	Gbps
• Scientific modelling	100s of Mbps - Gbps
• Intelligent data warehousing	Gbps and low latency
• Teleworking/distance education & training	6 –10 Mbps
• Desktop video-conferencing	6 –10 Mbps
• Publications	100s of Mbps
• Voice over IP	Not yet known
• E-commerce	Not yet known

V. THE HIGH-PERFORMANCE EUROPEAN BACKBONE NETWORK

The idea of a high-performance European backbone network has been promoted by many people and working parties, at numerous conferences and meetings (Williams, 1997; Axmann and Payr, 1997). While questions of topology and technical issues can safely be left to the experts, some general considerations are put forward in this article. They should serve not only to outline the user requests but should also give policy makers a better understanding of the needs and requirements of networking and the benefits to be derived from it.

The principles

The following is a first outline of the concept of a high-performance European backbone network. It requires further refinement and discussion. However, some of the basic principles are clear from the outset:

– *Quality, not only quantity:* High bandwidths for European research and educational networking are needed not only to satisfy the ever-growing volume of traditional networking services. New demands and applications, especially in the domain of multimedia-supported and real-time communication, including voice and video transfer, also require improvements in the quality of service offered by this new generation of research and education networks. The “best effort” approach of the existing Internet will remain the basis, but the network will also have to satisfy requirements for bandwidth-on-demand, low latency, multicasting, etc.

– *Production plus experimentation:* Although a high-performance European backbone network is primarily needed as a production network, it would be an ideal test-bed for new network protocols and applications. It does not seem unrealistic to assume that new network technologies, *e.g.* photonics, and the appropriate hardware will become mature within the next few years; what better test bed than a production network! The test bed must be appealing to academic and industrial researchers, the PNOs and the telecommunications industry alike, allowing them to run demonstrations and evaluations of new technologies in a pre-competitive "real" environment. The experience of the past has shown that, if appropriate precautions are taken, it is feasible to combine production and experimentation. Research and education environments, while pushing ahead with new applications and services, are an ideal test bed for pilot applications, ready to test and use "not-quite-perfect" solutions.

– *Applications follow availability:* The "Future of the Internet" advisory group stated some time ago that "... many useful and valuable applications will be developed in Europe, for both the commodity and high-performance Internets, as soon as bandwidth grows and pan-European coverage improves" (Williams, 1997). A concentration of European forces on advancing high-performance networking infrastructure in research and education could realistically create a fertile context for the development of business and employment opportunities in its applications and its migration to the commodity Internet.

– *Cohesion, not technology gaps:* A high-performance European backbone network as a joint European effort needs to take into account the rapid catching-up process in which some present and future member states of the European Union are involved. Therefore, the network architecture has to be sufficiently flexible to allow different bandwidths for links from national networks to the backbone infrastructure as well as provide for future extension.

– *European support for European added value*: The successful operation of a high-performance European backbone network depends as much on the efficiency of the overall management organisation as on the active participation of the NRENs. To guarantee the delivery of services to European researchers and scientists, whether in academia or in industry, and in support of Community activities, Community funds can justifiably be used to contribute to the overall running costs of the backbone network specifically for its European added value.

The benefits

In addition to the general and long-term effects of a high-performance networking infrastructure for research and education in Europe, some very concrete benefits can be expected:

– *Cost of intercontinental connectivity:* A joint high-performance European backbone network for research and education would strengthen the position of Europe in negotiations on intercontinental cost-sharing (Karounos, 1997).

– *Benefits from synergy:* The decision of the European countries to build a truly pervasive high-performance European backbone network that would secure the future of research, development and industry, rather than adding individual components in a piecemeal fashion, would certainly incur considerable costs in the short term. However, return on investment can be safely assumed to grow in correspondence with the synergetic effects resulting from co-ordinated development.

– *Academic-industrial partnerships:* A high-performance European backbone network should be attractive to the relevant industrial players on the European telecommunications scene. Academic-industrial partnerships are beneficial not only to research networking, but also to industry, especially as they pave the way for the large-scale commercial deployment of advanced communications infrastructure and technology, and will help to bring Europe fully into the Information Society.

VI. EFFECTS ON EUROPE

Universities and higher education institutions (in fact, educational institutions of all kinds), research establishments and cultural institutions use networks not only for their convenience; many have come to rely on them as a crucial instrument for communication. Networks provide the means for fast and efficient exchange of ideas, for setting up and maintaining collaborative work, distributing research results, offering teaching and learning materials to a wider audience, inviting review of ongoing work, producing works of art using new "materials" and new methods of presentation, and performing many other functions that would have been cumbersome to carry out in the "normal" way.

European research will benefit in many ways from a high-performance European backbone network: advanced networking applications, such as telecollaboration, virtual laboratories or video-conferences, help to keep up with the increasing speed with which research results now have to be produced and applied. The globalisation of research requires worldwide co-operation and sharing of resources. Co-operation is intensified and based on common interests and is no longer constrained by geography. Ever more sophisticated and costly research equipment can be shared and thus becomes accessible to all European researchers.

High-performance networking in education will play an important role in reshaping and redefining the educational system in response to the challenges posed by the Information Society: project-based, problem-oriented and resource-rich interactive learning will be made possible, independent of place and time, while maintaining the quality and intensity of teacher-pupil interaction. The sharing of courses and degree programmes among different sites and countries will ensure that students, wherever they are located, can obtain the best quality education possible.

The role of networking in research and education becomes even more evident when one looks at the global goals and priorities for European research and technological development (RTD) set out in the 5FP (European Commission, 1996) and when one tries to imagine whether any of these goals could be achieved without efficient and pervasive research networks:

- Confirming and strengthening the international role of European research by encouraging the co-operation of Europe's best research teams.
- Creating a research-friendly environment and maintaining a research context which is open to new ideas and which facilitates interfacing of basic and applied, of academic and industrial research.
- Creation of employment: Various studies show that the rapid and widespread introduction of advanced communications throughout the European Union will lead, directly and indirectly, to the creation of 6 million new jobs by the year 2010.
- Bringing research more into line with the real market as well as improving and accelerating the exploitation of research results; this is impossible without the efficient co-operation among partners from research, industry and user communities.
- Globalising research and strengthening partnerships throughout Europe so as to share the risks and pool the resources.

Community actions outside the RTD programmes will increasingly rely on and demand advanced networking. Examples in the current 4FP actions and programmes in the field of education and training include:

- SOCRATES in the domain of schools and higher education.
- LEONARDO in the domain of vocational training.
- The action line TMR (Training and Mobility of Researchers) with its focus on exchange of researchers and of access to research facilities.

Europe-wide co-operation among educational and research institutions plays an important part in all of these programmes. And while "physical exchange" (*e.g.* grant schemes for students and research trainees, international conferences and workshops, access to large-scale research facilities for researchers from all European countries) will certainly remain an important instrument for promoting European integration and Europe's scientific and innovative potential, "virtual exchange" in all its forms (such as virtual classrooms and laboratories, tele-conferences and tele-learning) will soon come to play an equally crucial role, bringing a new dimension and quality to European collaboration in research and education.

VII. CONCLUSION

The ideas, plans and actions mentioned above will not become relevant at some remote point of time in the future. The deadline for when high-performance European research networking should be on track for the future is already known. TEN-34 will soon be replaced by TEN-155, with support funding from the QUANTUM project of the Fourth Framework Programme. TEN-155 will also have a limited lifetime, but will at least ensure that support for European networking continues until the 5FP measures can come into effect.

Immediate action is therefore required, and the above-mentioned players need to contribute to the elaboration and implementation of a high-performance European backbone network. No one player can bring about the future of European networking for research and education on his own. Much work and many negotiations will be required, and co-ordination of all the various actors is of the utmost necessity.

NOTES

1. ENPG members (as of May 1998): Austria, Belgium, Finland, Hungary, Luxembourg, the Netherlands, Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, the United Kingdom. France, Ireland, and Italy have taken part in meetings as observers. TERENA has permanent observer status.
2. DANTE: http://www.dante.org.uk/.
3. Internet Society: http://www.isoc.org/internet/history/.
4. Internet2 Project: http://www.internet2.edu/.
5. University Corporation for Advanced Internet Development: http://www.ucaid.edu/abilene/.

REFERENCES

AXMANN, H.-P., P. LIZCANO and M. READ (1997),
"Report on Funding Research Infrastructure" delivered to the Telematics Programme Committee on 18 July 1997.

AXMANN, H.-P. and S. PAYR (1997),
"Towards a European High-performance Education and Research Network (HIPERnet)", JENC8, Session 412, May 1997.

EUROPEAN COMMISSION (1996),
"Inventing Tomorrow: Europe's Research at the Service of its People", Communication from the Commission. COM(96)332 final, Brussels, July 1996.

EUROPEAN COMMISSION (1997),
"Proposal for a European Parliament and Council Decision Concerning the 5th Framework Programme of the European Community for Research, Technological Development and Demonstration Activities (1998-2002)", COM(97)142, April 1997.

KAROUNOS, Theodoros (1997),
"Intercontinental Cost Sharing in the Internet between Europe and the USA", February 1997, http://www.dante.net/policy/co-fund.html.

TELEMATICS APPLICATION PROGRAMME (1996),
Report of the Five-year Review Panel, October.

WILLIAMS, David (1997),
"The Future of the Internet – What Role for Europe?" Interim Report of an Advisory Group, November 1997, http://nicewww.cern.ch/~davidw/public/i2eureport.htm.

COMPUTER NETWORKS AND THE VIRTUAL COLLEGE

Table of contents

This article was written by John P. Walsh of the Department of Sociology, University of Illinois at Chicago (e-mail: JWalsh@uic.edu) and Ann Roselle of Eastern Washington University.

OECD 1999

I. INTRODUCTION

While numerous publications have discussed the potential of computer networks to revolutionise scientific work, only a fairly limited number of empirical studies have attempted to measure the impact of this new technology. What follows is a summary of the existing literature on the effects of computer networks on scientific work and a discussion of the impact that this may have in the field of library and information science. Computer networks, which create a "virtual college" in this context, refer to computer-mediated communication (CMC) via a network (*i.e.* the Internet). Electronic discourse (e-mail, e-bulletin boards, e-conferencing, e-journals, e-chat) along with accessing computer databases and transferring electronic files are included under this broad use of the term "computer networks".

This article is organised as follows: Section II reports usage of CMC in select disciplines, while Section III gives details on remote collaboration, international collaboration, and changes in work group organisation brought on by CMC. Section IV explains levels of communication intensities facilitated by CMC, Section V introduces the idea of geographic co-operation through shared databases, and Section VI discusses changes in status distinctions and access to information due to CMC. Section VII introduces the effects of CMC on scientific productivity, Section VIII outlines social factors that may affect the use of CMC and, finally, Section IX summarises CMC use by and impact upon library and information professionals. Section X concludes.

II. USAGE ESTIMATES

Previous research on CMC suggests that the adoption of CMC technology can have a variety of effects on the social structure of work organisations (Sproull and Kiesler, 1991; Rice, 1994). Before a detailed discussion of these effects, it should be pointed out that they are tempered by usage, which varies significantly by field and by region.

While there has been little systematic, comparative analysis of usage rates of CMC across fields and across regions, estimates have been culled from various studies to get a sense of how usage may vary. For example, Walsh and Bayma (1996*a*), using membership listings, found that mathematicians (34% reporting e-mail addresses) and physicists (24%) made significant use of e-mail, even as early as 1991, while experimental biologists (9%) had yet to adopt this technology in large numbers (Table 1). Similarly, Hailman (1996) found that ornithologists in North America still had low rates (15%) of e-mail usage by 1993.

However, by 1993, the technology was beginning to diffuse rapidly. Based on a survey of aerospace engineers in industry, government and academia, Bishop (1994) reports that in 1993, 74% used networks, although only 50% had access to external research networks such as the Internet. Use among academics (rather than industry or government employees) and among scientists (rather than engineers or managers) was nearly universal (over 90%). Cohen's survey (1995) of chemists, sociologists, political scientists, and philosophers at Jesuit colleges and universities in the United States found significant differences in usage by field, ranging from 82% for chemists to 55% for philosophers (Table 1). Recent estimates by Walsh *et al.* (forthcoming) suggest that use has become widespread, but that there are still significant field differences, with usage estimates ranging from 78% for mathematics to 61% for sociology (Table 1). There is even variance across sub-fields of a given discipline (Walsh and Bayma, 1996*a*; Hurd and Weller, 1997). For example, Hurd and Weller (1997) compare chemists in different departments of a research university and find e-mail usage ranging from 84% for those in the Chemistry Department

Table 1. **Computer network use across scientific disciplines in the United States**

Field	Year	E-mail usage (percentage)	Year	E-mail usage (percentage)	Source
Experimental biology	1991	9	1997	70	1991: Walsh and Bayma (1996*a*)
Mathematics	1991	34	1997	78	1997: Walsh *et al.* (forthcoming)
Physics	1991	24	1997	66	
Sociology			1997	61	
All fields[1]	1992	39			Schauder (1994)
Ornithology	1993	15			Hailman (1996)
Aerospace engineering	1993	74			Bishop (1994)
Biological sciences	1993-94	89			Abels *et al.* (1996)
Engineering	1993-94	77			
Health sciences	1993-94	14			
Math./Computer science	1993-94	93			
Physical/Earth sciences	1993-94	80			
Social sciences	1993-94	55			
Chemistry	1994	82			Cohen (1995)
Sociology	1994	75			
Political science	1994	67			
Philosophy	1994	55			

Note: Numbers are not strictly comparable across studies due to differences in measures and sampling strategies.

1. Australia, United Kingdom, United States.

Source: Walsh and Bayma (1996*a*); Schauder (1994); Hailman (1996); Bishop (1994); Abels *et al.* (1996); Cohen (1995); Walsh *et al.* (forthcoming).

(College of Liberal Arts and Sciences) to 70% for those in the Department of Biological Chemistry (College of Medicine).

Thus, recent estimates suggest that CMC use has become quite common among scientists, highlighting the importance of the question of its impact on scientific work. However, even recent estimates still show significant variance across fields, suggesting that social or technical factors may influence use (Walsh and Bayma, 1996*a*). In addition, Hiltz and Johnson (1989) find that use (and satisfaction) do not necessarily correlate with beneficial outcomes. In the next sections, evidence that CMC is having measurable effects on scientific work will be explored.

III. COLLABORATION PATTERNS

Remote collaboration

One finding from the research on CMC is that it may lead to an increase in the size of one's social network. For example, in a study of oceanographers, Hesse *et al.* (1993) found that heavier e-mail users also report larger professional networks. Walsh and Bayma's (1996*b*) study of CMC use in biology, chemistry, mathematics and physics found that collaborations have been getting larger and that the effects seem to be associated with the use of this technology. As one experimental particle physicist explains:

> "Experiments are getting larger. We're trying to understand the nature of matter, the inside of a proton. We need more energy, more money and bigger equipment. There are fewer experiments now, because funds are limited. So, each experiment gets bigger. The size of experiments is growing. It's not a consequence of the nets. But the nets are a tool that made life easier. It hasn't affected the size of apparatus. But more people can collaborate effectively now." (Walsh and Bayma, 1996*b*)

One of the more significant changes in the work organisation of science has been the increase in remote collaboration, particularly international collaboration. Computer networks have been shown to reduce the need for co-workers to be co-located (Bullen and Bennett, 1991; Finholt and Sproull, 1990). Carley and Wendt (1991) claim that a new form of scientific work is emerging, what they call an "extended research group". These are very large, unified, cohesive and highly co-operative research groups that are geographically dispersed yet co-ordinated as though they were at one location and working under a single director. Such groups rely heavily on CMC to co-ordinate their work activities. Carley and Wendt (1991) point to the Soar group in artificial intelligence research as an example of this new form of working. Others have used the term "collaboratory" for a similar work arrangement (Lederberg and Uncapher, 1989; Finholt and Olson, 1996). Collaboratories provide the access to equipment, colleagues and databases that are traditionally part of the laboratory organisation of science, without regard to geography. Finholt and Olson (1996) list collaboratories in atmospheric and space science, biology, chemistry, medicine and physics. This new form of scientific work is exemplified by one of the early CMC collaborations, the Common LISP project. Orlikowski and Yates (1994) described the work group that developed Common LISP as a nation-wide network of computer scientists who communicated almost entirely via e-mail. A more recent example from computer science is the development of the Linux operating system, which has evolved through online collaboration of thousands of programmers across the globe (Harmon, 1999).

Empirical work shows a relation between use of the networks and collaboration patterns. Sanderson (1996) describes a collaboration in atmospheric physics that consisted of scientists at five Canadian sites, two US sites and two sites in another country. This research project adopted e-mail as its standard form of communication. All members of the research group have Internet addresses and most members of the group reported sending several e-mail messages per week to other members of the collaboration. E-mail was preferred over phone for the following reasons: scientists who travel may be hard to reach by phone, but can be contacted at their virtual address; written messages allow time to formulate answers before responding; and for non-English speaking colleagues, oral communication was more cumbersome than text. Sudweeks and Rafaeli (1996) give the history of ProjectH, an online collaboration of about 100 researchers in 15 countries to study CMC behaviour. The group planned the project, drew the sample (of e-mail messages on lists on BITNET, usenet and Compuserve), and developed the shared database without meeting face to face. The group also founded an online, refereed journal (*Journal of Computer-mediated Communication*), available on Web pages in Israel and the United States. Cohen (1995) found that there was a significant correlation between CMC use and co-authoring. While it is not clear whether CMC causes increased collaboration, results suggest that the technology certainly facilitates such collaborations (Walsh and Bayma, 1996*b*). For example, the paper that announced the discovery of the top quark listed 398 authors from 34 institutions in five countries (Walsh and Bayma, 1996*b*). The above quote from a particle physicist points to the importance of CMC in helping facilitate the very large collaborations in his field. Similarly, Bishop (1994) notes that aerospace engineers claim that one effect of computer networks has been to increase the feasibility and size of collaborative efforts. In addition, Walsh *et al.* (forthcoming) report that 65% of a sample of scientists in four fields report that e-mail use has increased their collaboration (Table 2). Furthermore, they report that e-mail use is positively associated with remote collaboration.

Table 2. **Percentage of respondents reporting a positive effect of e-mail on scientific/ professional outcomes, by field**

	Sample	Biology	Math.	Physics	Sociology	Field differences
Contact with scholars at other institutions	85.2%	86.8%	87.8%	77.5%	88.3%	**
Information about conferences	78.7%	70.0%	89.8%	81.9%	72.8%	***
Awareness of scholars at other institutions	66.0%	63.7%	77.6%	49.3%	72.5%	**
Amount of collaboration	65.4%	56.0%	63.3%	57.7%	48.5%	
Research productivity	58.5%	58.2%	69.4%	64.8%	41.7%	***
Awareness of calls for papers	58.0%	38.9%	71.4%	68.1%	52.9%	***
Distractions from research	50.4%	40.7%	44.9%	59.2%	57.3%	**
Information about grants	49.1%	51.1%	47.9%	45.8%	51.5%	
Breadth of research interests	41.9%	36.3%	51.0%	45.1%	35.0%	†
Academic prestige	20.3%	16.5%	28.6%	13.9%	21.8%	
Graduate student involvement in research	15.1%	17.0%	6.4%	18.6%	18.8%	
Ability to obtain grants	9.2%	6.7%	12.5%	9.7%	7.8%	
N =	314	91	49	72	103	

Note: Sample calculations have been weighted to control for field response differences.
*** p • .001 * p • .05.
** p • .01 † p •.10.
Source: Walsh *et al.* (forthcoming).

International collaboration

While science has always been an international activity, what has changed in recent years is the frequency of international collaborations. Table 3 shows the increase in the number of international collaborations since 1981, by field. The table shows that for all fields, the percentage of papers published with authors from more than one country has substantially increased. These collaborations are most common in fields that are more likely to use informal CMC, such as mathematics and physics (Walsh and Bayma, 1996*a*).

The scientists that Walsh and Bayma interviewed noted that their remote collaborations would be much more difficult, if not impossible, without the use of e-mail, with its low cost, speed and asynchronous nature. A respondent in experi-

Table 3. **Percentage of publications with international collaborators, worldwide and in the United States, by field, 1981-95**

Field	1981	1981-85	1986	1991	1991-95	1991-95	1995
	World	United States	World	World	World	United States	World
Math.	9	11	13	17	19	19	n.a.
Physics	9	14	11	16	19	25	n.a.
Biology	5	7	7	10	11	13	n.a.
Chemistry	5	9	6	9	11	14	n.a.
All fields	6	9	8	11	13	16	15

n.a. = not available.
Source: National Science Board (1993), Appendix Table 5-24; National Science Board (1998), Appendix Table 5-53.

mental solid state physics (working in the United States) noted the importance of networks for facilitating international collaboration:

> "Overall, it [networking] has been a gift. It's much easier to correspond and to get data. It's much more efficient. It's done a lot. I wouldn't be able to have these collaborations without networks. It would be impossible. We had two weeks' notice to put this proposal together with this guy in Austria. It would be next to impossible without the networks." (Walsh and Bayma, 1996*b*)

There was substantial agreement among respondents that one of the main virtues of this new technology is that it helps overcome some of the barriers created by geographic separation. This is most noticeable in international collaborations, where cost, time zones and language all create barriers that e-mail over the Internet is particularly well suited to overcome. This seems especially salient to researchers in Australia and New Zealand (Ostbye and Welby, 1988; Bruce, 1994). Bruce (1994) notes that one of the major benefits that his sample of Australian researchers attribute to use of the Internet is the ability to stay in contact with collaborators in the United States and Europe. Scientists in Australia and New Zealand have made heavy use of the Internet to obtain more ready access to research communities in North America and Europe. Table 4 shows the top-ten countries in terms of Internet hosts per 1 000 inhabitants. Australia and New Zealand both appear on the list, with net densities about the same as those for Canada or Sweden (Jacobson, 1994).

As a medium for facilitating collaboration, e-mail is seen as a next-best substitute for face-to-face interaction, and allows colleagues to continue to collaborate remotely after laying the groundwork in person (Carley and Wendt, 1991; Sanderson, 1996). Scientists emphasized the importance of establishing common understandings of the research problem through intensive, face-to-face interaction before engaging in computer-mediated collaborations. As one theoretical physicist explains:

> "It [e-mail] helps. But, you still need to talk and to work together. Once the project is in progress, once you finish the work, you can do the rest on line. It is difficult to do over the computer. You have to be close to the conclusion or

Table 4. **Number of Internet hosts per 1 000 inhabitants, 1996**

Rank	Country	Hosts per 1 000 inhabitants
1	Finland	62
2	Iceland	42
3	United States	31
4	Norway	30
5	Australia	24
6	Sweden	23
7	New Zealand	23
8	Canada	23
9	Switzerland	18
10	Singapore	18

Source: Ibrahim (1997).

> after. It's difficult to do research over the computer. You can communicate results. But it's not a substitute for discussion, the fighting that you have to do. In September, my collaborator is coming from Brazil. We'll do a lot of work while he's here. We'll get the essence of the paper together. Then, we'll add stuff later [via e-mail]." (Walsh and Bayma, 1996*b*)

While science has always been a fairly international endeavour, this globalisation of fields is creating tighter links among scientists who are geographically dispersed, leading to a more closely knit international community of scholars. CMC has not caused these changes. It has, however, provided the infrastructure that allowed them to occur. International collaboration was always possible, but it was fairly rare before networks were in common use. Table 3 shows that, even now, although more common than ten years ago, international collaboration is still comparatively rare in the fields where informal CMC is less common, such as chemistry and experimental biology (Walsh and Bayma, 1996*a*).

Reorganisation of collaborations

This change in collaboration patterns suggests that collaborations are becoming mediated more by substantive fit, rather than by geographic or personal linkages (Van Alstyne and Brynjolfsson, 1996*a*). Previous research on CMC (Finholt and Sproull, 1990; Feldman, 1987) found that CMC leads to a reorganisation of work groups emphasizing substantive overlap and discounting geographic overlap. Such technology allows the creation of work groups that are larger, more complex and have more fluid structures than face-to-face communication can support (Rice, 1994, Chapter 8). Like the virtual corporation, networks help create virtual research teams which link a variety of scientists, with each contributing his or her unique skills to the project. Projects can take advantage of the networks to access exactly the skills that are needed. Researchers can take advantage of the networks to gain access to a wider variety of projects that can make use of their skills. Consistent with that result, scientific fields have been changing their work organisations, and respondents attribute these changes in part to the existence of CMC. A common theme in each situation is that CMC is adopted in a way that reproduces local social relations and research practices among remote collaborators. Thus, while there is some reorganisation of the social structure due to CMC use, this reorganisation seems to be largely limited to changing (expanding) who can participate, with only minor changes in the content of participation in the research group. For the most part, the existing work organisation has been reproduced over a wider geographic area – with CMC serving as the link that was formerly served by face-to-face communication in local collaborations (Allen, 1977).

This change in work group organisation could become one of the most revolutionary impacts of this new technology. By removing the geographic component of work structures, computer networks are facilitating a wide variety of new work arrangements, including multi-site work teams, telecommuting, distributed organisations, virtual organisations, and other forms of geographically dispersed but networked work teams. However, there is some debate about whether this change is always beneficial for those affected. For example, telecommuting allows workers to work at home and saves them the hassles of commuting and the cost and time required to put together a professional presentation of self each morning. At the same time, it isolates workers from the social and occupational stimulation that comes from informal interaction with co-workers. Similarly, if workers begin to form work groups with remote colleagues, their face-to-face interactions with their local colleagues may be replaced by computer-mediated interactions with remote colleagues (Van Alstyne and Brynjolfsson, 1996*b*). While CMC may facilitate cross-disciplinary collaboration, it can also lead to a balkanisation of science, with researchers using their limited communication time to interact only with those in their speciality (anywhere in the world). It may be the case that this balkanisation of scientific communities will reduce the cross-fertilisation of ideas from one discipline to another. The impact of these new forms of scientific work organisation on scientific outcomes is not yet clear.

Thus, e-mail is seen as a way of extending one's network of potential collaborators and facilitating existing collaborations by allowing personal contacts to continue as collaborations even after the colleagues have separated. CMC is not, however, seen as a complete substitute for face-to-face (often involving travel) or telephone interaction.

IV. FREQUENCY OF COMMUNICATION

CMC technology, compared to face-to-face interaction, is also less constrained in terms of the total amount of communication generated. In a face-to-face meeting, generally only one person can talk at a time and the total time for talking is usually strictly limited. Some types of CMC, on the other hand, allow all participants to talk at once. The discussion can also continue for a longer period, because all do not have to be simultaneously brought together to the exclusion of other activities (Siegel *et al.*, 1986; McGuire *et al.*, 1987; Dubrovsky *et al.*, 1991; Weisband, 1992).

In academic science, we find evidence that CMC may contribute to an overall increase in the amount of communication performed during a research project (Walsh and Bayma, 1996*b*). In high energy physics experiments, e-mail has allowed researchers to maintain involvement in long-term experiments even when not physically co-present at the lab, through the use of distribution lists, bulletin boards and e-mail and by distributing pre-prints and other crucial (informal) information electronically. This method of communication allows all members of the collaboration to stay "in the loop". Because memos can be distributed simultaneously to all members, those who are not physically located at the lab can participate in the ongoing discussions of the experiment (Knorr Cetina, forthcoming). However, CMC itself is not sufficient. "Computer-mediated communication" represents a variety of communication genres (Orlikowski and Yates, 1994). The selection of those that are to be used affects the outcomes. Also, the ways in which the particular technology is used are important.

An additional barrier to communication in scientific work is time dispersion. Because physics experiments often run on three shifts, e-mail provides a convenient way to keep all shifts informed. Previous work on police organisations has shown how shift work interferes with information flows (Maltz *et al.*, 1991). Previous work on electronic mail has shown how e-mail can help overcome some of the informational and social psychological isolation of shift work (Huff *et al.*, 1989). CMC can play an important role in integrating work teams that are both geographically and chronologically dispersed. A solid state physicist points to the importance of CMC in his collaborations, contrasting CMC with previous forms of co-ordination:

> "We worked together [in collaboration] before the nets. We used different modes. We would visit each other. The nets allow an intensive means of collaboration. You can communicate as if you were next door. If you want to you can call up data, look at it on the screen. It's as if you are together, far away. The time difference can even be an advantage. I send a message at the end of the day. He works on it while I'm asleep. I get the answer in the morning my time." (Walsh and Bayma, 1996*b*)

This suggests that CMC allows remote collaborations to be more communication intensive than might otherwise be the case. Rather than extended isolation punctuated by periodic, intensive face-to-face meetings, CMC collaborations can maintain a high level of contact among the remote collaborators. Also, CMC helps overcome chronological dispersion (in this case, created by time zones, rather than shift work). Furthermore, chronological dispersion is now seen as an advantage. By passing the research tasks back and forth, collaborators can create the project that never sleeps (Sudweeks and Rafaeli, 1996).

Similarly, chemists and biologists noted that the networks allow closer communication with remote colleagues. This is particularly important for chemists and biologists whose loosely coupled experiments would often become out of synch as a result of the infrequency of traditional forms of communication. As one chemist explains:

> "We run projects with lasers, high-vacuum technology, ion beams, molecular beams. It takes years to put these things together, years to do the experiments. Things have changed a lot with the advent of commonly usable e-mail. You'd have an experiment in California or somewhere, maybe every few months

> you'd call and co-ordinate with them, so inevitably there'd be problems. Now, with e-mail, that's no problem." (Walsh and Bayma, 1996*b*)

One effect of this increased density of communication may be an increased attachment to the research group and the discipline. Scientific work can be very alienating, in part due to the isolation that comes from irregular hours and from concentration on a highly specialised endeavour that even local colleagues may have little interest in. As Huff *et al.* (1989) note, electronic mail use can help overcome the sense of isolation and thus lead to increased commitment. The following quote by a physical chemist who uses the networks for international collaboration points to one of the important features of the network – its ability to keep weak ties active:

> "Gossip – definitely – it's a very important component of messages. If you only see people once a year, you need to let them know you care about them as a person, not just as a disembodied voice at the other end of a phone line. Otherwise, I don't think you could maintain a long-distance project like this." (Hurd and Weller, 1997)

If one thinks of a network as a set of connections which decay over time unless occasionally activated (somewhat like the way in which neural networks are generally conceived), then electronic communication provides an important mechanism for maintaining ties with others who would probably get lost from the network if communication depended on more formal or intrusive means. A common use of the networks is to send an occasional note whose latent message is "I know you are around and I'm still interested in maintaining a tie with you". Table 2 shows that the largest effect of e-mail is to increase contact with those at other institutions. Sproull and Kiesler (1991, p. 101) conclude that, while not a complete substitute for face-to-face interactions, CMC can provide an avenue for social interaction for what might otherwise be isolated individuals. This increased interaction can lead to increased job satisfaction and commitment.

On the other hand, it has been noted that e-mail communication does away with many of the pleasantries and socialising that generally come with face-to-face or even phone conversations. While some respondents viewed this as an advantage of e-mail (in the sense that time costs are reduced), this also creates a more instrumental and barren collegiality.[1] While CMC may allow more frequent and interactive communication among remote colleagues, it can also create a more distanced, instrumental communication structure among colleagues. This effect can spread to those within the same institution. Scientists have noted that they often used e-mail to communicate with local colleagues, citing the advantages of asynchronicity and less time wasted in pleasantries as reasons to choose this mode. While this may allow messages to move more quickly, it could also create a more socially isolated and alienating work environment. CMC may, in fact, be simultaneously integrating and isolating individuals. CMC allows more frequent and individualised contact with others who are remote (in terms of geographic or social distance). At the same time, it may make those interactions more barren. The result could be a work environment, or community, where each is linked to more others, but where those links become more instrumental or less satisfying. Further research is needed to specify the conditions under which CMC becomes a medium for integrating dispersed individuals into virtual groups *vs.* CMC being a medium for isolating and alienating individuals (Walther, 1996).

V. GLOBAL SCIENCE

In addition to its effects on collaboration, CMC may also be increasing geographic co-operation of a less explicit nature (Sanderson, 1996). Scientists in a variety of fields are producing data sets that are accessible via the Internet to scientists across the globe. Perhaps the most famous examples come from biology. These include the Human Genome Project and the Protein Data Bank, which provide biologists with a growing data set of information on genetic material and proteins (Walsh and Bayma, 1996*a*). These data sets (which take online contributions and which are accessible on line) provide scientists across the globe with mechanisms for sharing data and combining the results of loosely co-ordinated studies. In fact, many leading journals now require data to be deposited in these databases as a condition of publication. This global science is not limited to biology. The Internet is providing new opportunities for scientists in different countries to combine local data sets and create global data sets that can be used to answer questions that depend on collecting data from across the globe. For example, several projects combine global geoscience data (Bierly, 1988). Sanderson (1996) notes how the atmospheric physics group studied used ftp and remote log-on to both maintain the integrity of the very large file of cleaned satellite data and allow research groups in three countries to access the data. In the social sciences, Rainwater and Smeeding (1988) describe the Luxembourg Income Study, which currently has data on 20 countries. This project allows researchers to conduct comparative analysis of individual-level data by sending SAS or SPSS programs via e-mail to the data centre in Luxembourg. The data centre acts as a repository for the data, and also screens the programs to ensure data confidentiality. Thus, researchers can access, at no cost, this micro-level data from 20 countries,[2] without having to travel to Luxembourg, while maintaining the confidentiality that is necessary for the collection of such data. In the humanities, Ruhleder (1994) and Everhart (1996) describe similar uses of the Internet in the humanities, where the data sets are texts of classical or medieval literature. Thus, the Internet may be creating a global science, both in the sense of international collaboration discussed above, and in the sense of integrating data from across the globe and making it readily accessible to researchers, independent of location.

VI. PERIPHERALITY EFFECTS

Status distinctions and access to information

In general, there is still significant dissent about the impact of CMC on work organisations, particularly with respect to peripheral effects. Several studies have found that CMC can lead to more decentralisation or reduced status differences in an organisation or group (Dubrovsky *et al.*, 1991; Sproull and Kiesler, 1986; Rice, 1980). In general, this finding is attributed to the fact that interaction on the Internet provides fewer status cues than does face-to-face communication, or even mail or telephone interactions. Important distinctions that are removed include differences of rank and gender. Group decisions are less influenced by the status of those proposing particular solutions (McGuire *et al.*, 1987). However, to the extent that status distinctions are reproduced in this new communication genre, high status individuals will continue to exert more influence on group decisions (Orlikowski and Yates, 1994). Weisband *et al.* (1995) found that experiments comparing CMC and face-to-face groups showed little difference in the inequality of participation in decision making. They suggest that one explanation for the difference between their results and previous work is that group members were aware of the status distinctions among members even in the CMC condition. As the technology becomes more developed, it is beginning to insert more status cues into the communication, such as more decipherable e-mail addresses and pictures and biographies on Web pages. Also, other mechanisms for reintroducing the status-reinforcing procedures that characterise previous communication technology (such as mail and telephone) are beginning to emerge. For example, high-status scientists (and presumably high-status managers as well) will use gatekeepers to screen their e-mail in the same way they screen calls and letters. Similarly, to the extent that such technology violates existing work norms or status distinctions, the technology may not be used and hence will have little impact (Sproull and Kiesler, 1991; Orlikowski, 1993).

In addition, new technology can transform some of the bases for existing status distinctions. Barley (1990), for example, shows that technological changes at work can enhance the status of younger colleagues because of their greater familiarity with the latest technology. Computer networks may also provide wider access to crucial resources, such as computing facilities, software or databases (Dongarra and Grosse, 1987). In the past, access to resources was unequally distributed in science (Merton, 1968; Cole and Cole, 1967; Allison and Stewart, 1974). Because networks can make these resources available to more scientists, they could increase the productivity of these peripheral scientists (*i.e.* those less senior, less eminent or not located in major institutions). This greater access could reduce the gaps between the core of the field and some of their less eminent colleagues. A recent study on network use in oceanography found that younger oceanographers who were heavier users of the networks were more likely to receive professional recognition than were age peers who did not use the network so heavily (National Science Board, 1993). Similarly, inland oceanographers who used the networks more had more publications than those who were less active on the networks. These results suggest that the technology may be transforming the social structure of oceanography by providing greater access for lower status scientists. However, Cohen's survey (1995) finds no evidence of democratisation effects. For example, there was no added advantage of CMC for those at less prestigious institutions, nor for younger scientists.

One outcome of the net's ability to maintain ties with remote collaborators (as noted above) is that scientists who were trained at research centres can maintain their research contacts when they move to more peripheral institutions. One mathematician, when questioned about the need to be on the net in order to be research-active states:

> "There's a lot of people here, lots of people who visit here, who do my kind of work. I can do research without the net. But, I'm in a temporary position. Next year is my last year. Where I go next, it might be true that you have to be on the network. For example, if I end up in New Mexico State in Las Cruzas, e-mail may be crucial to keeping my research activity going." (Walsh and Bayma, 1996*b*)

Again, this result suggests that electronic networks are not necessarily replacing face-to-face contact or eliminating the importance of research centres in these disciplines. Rather, they are being used to keep the connections built up during these contacts from atrophying. In this sense, the set of active researchers in a field may be expanding, because those who move to more peripheral institutions are better able to maintain the ties needed to stay research-active. In a similar vein, one biologist argued that national and international conferences have risen in importance, now that it is possible to establish familiarity with dispersed participants' work and worthwhile to make remote ties.

As the following experimental physicist notes, this peripheral effect generally takes the form of increased participation by less prestigious institutions, rather than in a restructuring of the status hierarchy:

> "I could say they [smaller/less prestigious institutions] get exposure to what's going on at the big institutions, but I can't think of an example of a less prestigious institution in Italy [his home country] gaining anything in terms of prestige from the net." (Walsh and Bayma, 1996*b*)

Similarly, a mathematician notes that the changes attributed to the network are largely due to including more researchers at the bottom rather than transforming the work of those at the top:

> "Really creative math is an individual creative process. There is a lot of information available. A lot of research is not earthshaking. A lot, some of the big stuff, is working out of big things. Real creative stuff. The networks aren't that

> big a factor. If you are in communication with the giants of a subject, then you have a better chance of getting in on new ideas early on. You should be part of that transference of information. Some of the strongest mathematicians in the world don't have anything to do with computers. For mathematicians of average ability, e-mail enhances their opportunity to, as the Army says, 'Be all you can be'. Here at the top, they can't be much better, and no one would dare pressure them to change. E-mail helps. If you're sentenced to Podunk, wherever that is, it's not the death sentence it used to be." (Walsh and Bayma, 1996*b*)

As the above quotes point out, e-mail does allow those who would previously have been excluded from the most up-to-date information to access that information. It is claimed that such access is important to staying current in the field and those at less prestigious institutions are now able to participate in the scientific communities. Bruce (1994) notes that his sample of Australian researchers points out that the net allows them to stay current in their fields. CMC also facilitates access to other types of resources besides potential personal contacts, such as shared databases or computing facilities (Walsh and Bayma, 1996*b*; Finholt *et al.*, 1995). However, these respondents also suggest that the result is not a levelling of science so much as an expansion of science. Those individuals and institutions at the top continue to be the central figures. The difference is that now there are a wider variety of individuals and institutions who are contributing to the body of scientific knowledge, or at least who are able to more closely follow the advances in their field.

An additional example of this effect of expansion rather than levelling comes from theoretical physics. Physicists (particularly, particle physics theorists) have begun to institutionalise the electronic distribution of pre-prints. One important channel of informal communication in science is pre-prints. In the 1970s, Garvey and Griffith (1979) noted that the distribution of pre-prints had been increasing to the point where it could become quite burdensome for authors. They also noted that "those who need pre-prints most – young scientists, workers at small institutions and researchers in less developed countries – are frequently not the recipients." (Garvey and Griffith, 1979, p. 158). In the early 1990s, a theoretical physicist at Los Alamos set up an electronic pre-print bulletin board (Taubes, 1993; Merz, forthcoming). This bulletin board provides subscribers with abstracts of all the new papers and allows subscribers to send an e-mail request and receive the full paper, with graphics, electronically. The service began in August 1991, and by December 1992 had 8 000 subscribers. The database was soon receiving about 600 new papers per month.

The pre-print database also provides an example of how importing genres of communication from one medium (paper pre-prints sent by post) to another (an electronic bulletin board) is a change that both transforms and reinforces the existing structure of communication within a community (Orlikowski and Yates, 1994). This pre-print database has contributed to the inclusion of a greater number of theoretical physicists worldwide into the pre-print loop. However, as the following theoretical physicist points out, this peripheral effect is tempered by the need for access to local information to process the papers in the pre-print database:

> "They've [peripheral institutions] benefited. If you're in the Third World, the underdeveloped countries or at a small institution, in principle, access is easy now. There are barriers still. Before, you were cut off because you couldn't get the information. Now, in principle, someone in India can get the information the same as someone at Princeton. That's a big change. However, it's different to see a paper and to be there. If you are at the big institutions you have access to the oral information, seminars, you can talk to the person. That's still lacking. They're getting in over 100 papers per month. At the big institutions, you have someone who knows what is important and you can get help sorting through it. The networks help those with the necessary expertise. They can profit from this new access." (Walsh and Bayma, 1996*b*)

Garvey (1979) has argued that the filtering and pointing provided by this local, informal communication is an important part of the process of scientific information searches. The above quote suggests that, while CMC helps, by itself it is not sufficient to overcome the disadvantages that come from not having direct contact with the most active scholars in one's field. While the pre-print database, in effect, lets a wider group of people listen in on the paper dialogue among the research leaders in the field, only a small sub-set are positioned to take advantage of the quicker and wider access to information the database provides. Most of those who make up that sub-set are the research leaders who were already exchanging papers and communicating through other informal means.

E-mail also facilitates establishing ties that might otherwise not get created. The following quote from a mathematician reflects how CMC frees communication from some of the social inhibitions of other forms of communication:

> "Once you open communication you can have interactions with people. I might not walk into Harvard for a face-to-face meeting, but I can send e-mail to Harvard. With my son, he would get on and contact prestigious people in his field. He didn't know who they were, but I recognised some of the names. You might not go face to face, but on e-mail, you don't know who you're talking to." (Walsh and Bayma, 1996*b*).

The following quote, from a theoretical physicist, demonstrates the same point from the receiving end of such communications:

> "I haven't initiated any work with strangers via the net. But, I have been on the receiving end of such requests. Others have written, 'This looks interesting. We're working on X. Can you give us some suggestions'. When they get a result,

> they ask to publish it jointly. It used to happen with letters. But, the informal nature of e-mail helps. You tend to think more about a letter. [E-mail] is more informal. It helps. You can send a short message where with a letter, you might not send it." (Walsh and Bayma, 1996*b*)

As Sproull and Kiesler (1986) note, the lower social context cues and the informal nature of e-mail reduce the constraints on lower-level individuals contacting higher-level individuals. The above quote suggests that scientists who might be reluctant to send a letter to generate a contact (because of status differences, for example) would be willing to initiate contact via e-mail. Thus, e-mail may be facilitating the creation of new ties between remote collaborators and give lower status scientists the ability to query their more eminent colleagues. The impact of this potential change is not certain. The quote above suggests that such contact might lead to even more disparity in publication rates, as the top scientists become attached to an ever greater network of research projects generated through e-mail contacts. Alternatively, such contact may allow scientists who previously lacked the access needed to stay current to become active participants and perhaps future core members of their fields. Note, however, that Cohen (1995) found no significant interaction between age or institutional prestige and CMC use as a predictor of productivity.

Networks may also be breaking down the strong correlation between distance and informal queries (McGuire *et al.*, 1987; Kraut *et al.*, 1990). Among scientists, the ability of the networks to quickly disseminate questions to a large number of people through electronic discussion groups allows scientists to interact in a "down the hall" way with people spread over a large area (Sanderson, 1996). The following quote from a mathematician is a typical example:

> "I used [sci.math] once. I wrote a paper and sent it to the editor. He told me: 'That's related to X result of so and so.' But he didn't have the reference. So, I posted the question on netnews. 'Can anyone give me the reference?' Within a day, two or three people sent the precise reference. It's an easy question, but it's hard to dig out of the library. It would take hours. I can ask several thousand people at once. There's a chance that two or three know. It doesn't happen otherwise. There's no way [besides e-mail] to ask an insignificant question of a few thousand people at once. It's important to me, but not to anyone else. They were happy to help out." (Walsh and Bayma, 1996*a*)

Constant *et al.* (1996) find that CMC facilitates information exchange with those in other units of a large multinational computer company. Finholt (1993) finds, in the same company, that CMC allows access to organisational memory through its database of past broadcast queries. Table 2 above shows that the biggest effect of e-mail is increasing contact with those at other institutions. However, Van Alstyne and Brynjolfsson (1996*b*) argue (based on a formal, economic model) that this increased access to colleagues, independent of geography, when combined with economising on information exchange and maximising benefits from exchange partners, will lead to increased inequality in science, as each person limits his interactions to his information peers, independent of location. Further empirical work is needed to sort out these contradictory findings and predictions.

VII. PRODUCTIVITY

Computer networks may also increase the productivity of scientists, although here the evidence is generally correlational, rather than causal. Several previous studies have found that those who make the most use of computer networks also tend to be the most productive. However, it could be the case that those who are most productive are the ones who have the most need or the most interest in taking advantage of the benefits of CMC. One early study of computer conferencing systems (including EIES in the United States and COM in Sweden) found that objective measures of usage were significantly positively associated with subjective measures of productivity and career advancement for samples of scientists and R&D personnel (Hiltz and Johnson, 1989; Hilzt, 1988). Hesse *et al.* (1993) found a positive correlation between computer network use and productivity, even after controlling for other variables related to productivity, such as age and prestige of institution. In a study of academic chemists in the United Kingdom, Philip (1995) found a correlation between the ranking of the department and the use of online chemical information systems. Cohen (1995) also found a significant correlation between CMC use and publications and professional recognition, again after controlling for other related variables. Kaminer and Braunstien (1998) have also found this positive productivity-CMC relationship.

There is some evidence that users attribute productivity effects to CMC. Cohen's respondents associated greater productivity with the benefits provided by CMC (particularly, easier communication with colleagues at other campuses). Similarly, academics in Australia noted that the networks increased their work efficiency, although this study did not measure productivity directly (Bruce, 1994). Bishop (1994) found that, among aerospace engineers, respondents felt that networks increased the amount of information available, enhanced efficiency in contacting people, and increased their ability to complete projects on schedule, suggesting that networks do improve worker efficiency. However, her study also found that networks might increase "the amount of time spent fooling around". Walsh *et al.* (forthcoming) find that more than half of their respondents felt that e-mail led to higher productivity, and that mathematicians and physicists were especially likely to report a productivity effect (Table 2). They also find a positive relationship between CMC use and papers published. However, like Bishop, they find that a significant number of respondents felt that e-mail also increased distractions from research (Table 2).

There is even less evidence of the impact of the Internet on the productivity of science (as opposed to the productivity of scientists). One could argue that the quicker accumulation of articles and the access to more data should allow scientific discoveries to follow more quickly one after the other. Merz (forthcoming) suggests that the pre-print archive in physics may increase scientific productivity, not only by making research more efficient and accelerating publication, but also by synchronising research activity and allowing a more efficient division of labour across research teams. Uzumeri and Snyder (1996) argue for an acceleration of science due to the Internet, in this case, regarding the discovery of product defects. They compare the case of the Dalkon Shield's flaw, which remained secret for six years, and the Pentium chip's flaw, which was revealed in six weeks as a result of activity on the Internet. Still, the question of the impact of the Internet on the pace of scientific discovery has not been answered. Cole and Cole (1972) argue that the pace of science is driven primarily by those at the top of their field and that reducing the participation of the majority of scientists would have little impact. By extension, their argument (the Ortega hypothesis) suggests that expanding the participation of those at peripheral institutions should have little impact on the pace of science. While difficult to measure and test, this question of the impact of this new technology on scientific progress is the key question in need of an answer.

VIII. SOCIAL CONTEXTS

Many of the effects of CMC may be contingent on various social factors (Kling, 1995). While such technology can be used to democratise access to various resources, it can also be used in a way that ensures that only those at elite institutions have access to such resources (Rice, 1994). Orlikowski and Yates (1994) argue that the genre repertoire for communicating in an online community is heavily influenced by members' standards of interaction brought from previous modes of interacting. While there is some evidence of the impact of CMC, we suspect that the context in which CMC is introduced is an important mediating factor in explaining the effects of CMC (Sproull and Kiesler, 1991). For example, Walther (1997) finds that the structuring of the collaboration and respondents' attitudes towards that collaboration (long term *vs.* short term and group *vs.* individual) interact with the effects of CMC, such that some conditions produce greater effort than face-to-face groups, and some produce less effort.

Peripheral effects depend on universal access. In many settings, such assumptions are not met. For example, while access to the Internet among US academics is nearly universal (Jacobson, 1994; Gurbaxani, 1990), access is much less uniform outside the United States (Jacobson, 1994). In addition, CMC may be structured in such a way as to reinforce status distinctions. For example, netnews bulletin boards, such as the sci.* hierarchy, can be used to link the members of a field. These bulletin boards can be used to announce new findings, to discuss substantive issues and to obtain answers to questions from unknown colleagues. A similar mechanism is the field-specific distribution list. However, one important difference is that the existence of a netnews bulletin board is broadcast to a large number of computer servers ("the world"), while the existence of the distribution list may be announced only through direct contact through existing research ties. The list may even be limited, so that only a select group can subscribe. By adopting the first form, anyone with netnews access can find out about and participate in the discussions of the electronic group. By adopting the second form, only those select few will have access to the information posted to the list. While both cases are examples of CMC, the outcomes in terms of access to those not already directly connected to the core scientists in a field may be quite different.

In addition, there is some evidence of a fit effect. Some fields seem to have greater potential to benefit from this technology than do others. Walsh and Bayma (1996*a*) and Bishop (1994) suggest that fields where interdependence is high (*i.e.* where collaborators need to react back and forth to each other's activities on a frequent basis) and where collaborators are likely to be dispersed (mathematics, physics and aerospace engineering are examples) are the fields most likely to benefit from the ability of CMC to provide reliable, quick, asynchronous communication. Hailman (1996) suggests that low usage in ornithology can be explained in part by technical limitations (the difficulty of transmitting the non-text information that is common in this field) and the relatively slow pace of discovery (where old literature is still relevant and publication lags are not considered problematically long).

One impact of computer networks is that they make information more widely available. While this has several advantages, as noted above, there is also a negative interpretation of this effect. Walsh and Bayma (1996*a*) note that many scientists were reluctant to use computer networks because they might lose their ownership of valuable information. This was particularly salient for those in chemistry and experimental biology, where there is a significant industrial presence and where patent rights may be worth substantial sums. Similarly, Bishop (1994) found that aerospace engineers (many of whom do defence work) were quite worried about the effects of computer networks on system security and leaks of proprietary information. This problem of information ownership will continue to plague policy makers trying to increase access to and utility of the Internet while protecting property rights over information.

IX. IMPLICATIONS FOR LIBRARY AND INFORMATION SCIENCE

Library research and CMC

Although there is a significant amount of literature within the field of library and information science reflecting the use of CMC for day-to-day practical tasks and responsibilities, literature examining the effect of CMC on *research activities* within the field is more scarce. This fact could possibly be correlated with the smaller amount of scholarly research being conducted in library science compared to those science disciplines already discussed.

Nevertheless, Kovacs *et al.* (1995), in a national study on the use of e-conferences by library and information science professionals, discovered that this electronic means of scholarly communication was serving a role in research. A majority of respondents in the study indicated that e-conferences enhanced other sources of scholarly information, such as professional journals, professional conferences, telephone conversations and postal mail. Even more so, 38.4% reported that e-conferences were replacing these other forms of traditional communication. Finally, 33% of the respondents indicated that they participated in e-conferences for professional and research interests. Along with the over 130 electronic discussion lists for librarians (Abbott, 1994), refereed e-journals which disseminate research results are beginning to emerge, including *LIBRES, Library and Information Science Research Electronic Journal*; *MC Journal; The Journal of Academic Media*; and *Public-Access Computer Systems Review.*

In 1991, the Association of College and Research Libraries (ACRL) established a mentoring programme via e-mail with the purpose of encouraging and improving research in library science (Echavarria *et al.*, 1995). ACRL recognised that the pool of active researchers was small and that CMC would provide the mechanism to reach a larger number of professionals with the hopes of encouraging research. Although the programme has had some failures, it also has had lasting impacts on participants' research skills. As one protégé explains:

> "This summer I transferred from the cataloguing department to the reference department, and this decision to change course was motivated by being part of the [e-mail mentor programme]. I realised that the environment in the cataloguing department wasn't conducive to doing research, and I eagerly sought contact with library users. Being a part of the [e-mail mentor programme] has had a strong influence on what I have read, what I have thought and where I find myself today." (Echavarria *et al.*, 1995, p. 358)

Another innovative use of CMC is the establishment of virtual programmes for obtaining a master's degree in library and information science. University programmes at the forefront of Web distance education include those at the University of Arizona, the University of Illinois (Urbana-Champaign), and Syracuse University.

The University of Arizona virtual programme, which in 1997 had approximately 70 students worldwide, uses a combination of CMC. Lectures are posted over the Internet via a Web page, class discussions occur via an electronic discussion list, assignments and tests are distributed via ftp, student group projects are co-ordinated through e-mail and final projects posted on a Web page, and students communicate with their professors via e-mail. It should also be pointed out that students are required to take 12 credits at the University of Arizona campus. As noted above, face-to-face interaction is important in collaborative effort. It will be interesting to observe what lasting affects this virtual training will have on these new library and information specialists, and whether or not they will be more inclined to collaborate with others via CMC when they engage in research.

Role of the librarian in the virtual college

Librarians and information specialists have always paid special attention to the ways in which scientific knowledge is created and disseminated. Keeping track of the virtual college during these processes is an additional challenge for librarians. During the University Libraries Section of the 1997 Annual Conference of the American Library Association, a discussion of this issue was presented by both librarians and academic scholars (Garrison *et al.*, 1997). One librarian suggested that as scholars change in their methods, so too will librarians need to change. Likewise, an academic scholar indicated that librarians may be moving from being keepers of collections to becoming finders and organisers of information (Garrison *et al.*, p. 544).

As scientific work and communication increasingly occurs within an electronic arena, signs of librarians proactively playing an integral role in finding CMC resources and organising electronic information to enable the majority to access and make meaning out of these resources are beginning to appear. The Association of Research Libraries publishes a comprehensive directory of electronic journals, newsletters and academic discussion lists. Individual librarians also create bibliographies and directories to CMC resources (Abbott, 1994). In addition, librarians attempt to keep track of scholarly information available over the Internet by providing in, an organised manner, links to such information via library home pages. These links are made in order to improve access for individuals previously excluded from such networks of information.

Finally, perhaps, now that a virtual college exists, librarians (and especially research or academic librarians) need to rethink how they are relating to scholars. Librarians are developing a virtual community on their own while separate virtual communication is occurring for scholars within various disciplines. It might be valuable to question whether or not librarians' communication patterns with scholars need adjustment. There could be benefits, for both parties, in the overlapping within these virtual communities. For example, it has been argued that librarians, through instruction and close contact with faculty, can assist in preventing the balkanisation of the scholarly communication network which is often caused by variation in levels of faculty technological expertise (Schwartz, 1994).

X. CONCLUSIONS

As CMC has become more salient to academic work, there has been significant growth in research on the impact of this new technology on science. Prior empirical work on the effects of the Internet on science suggests that scientific work is changing in profound ways. The most significant change may be the transformation of collaboration patterns. There has been a growth in remote collaboration and, in particular, a growth in international collaboration. These changes have not been caused solely by the presence of this new technology. Rather, the new technology has facilitated this change in work patterns, by allowing researchers ready access to like-minded colleagues who happen to be at another institution (perhaps many time zones away). Yet, several studies point out that these collaborations are often punctuated by face-to-face meetings, to establish a common understanding of the research question and procedures and to allow for mid-course corrections. Longitudinal data and field experiments to test for the link between CMC use and changing collaboration patterns are still needed, separating out the effects of other changes in scientific problems and institutional conditions (such as funding policies) to determine what the actual impact has been.

The question of the impact on the status hierarchy in science is much more in doubt. Some studies have found that the Internet provides new opportunities (by providing access to data, instruments, colleagues, results, etc.) for those at early stages in their career or at more peripheral institutions (either in terms of geography or status). Others have found little evidence of differential benefits for those outside the centres of science. It seems that the major effect has been to increase the number of researchers and institutions that can participate in the dialogue of a scientific community. Whether this participation consists primarily of listening in on the conversation or making significant contributions that might otherwise have been missed still needs to be demonstrated.

Finally, there is some evidence that CMC improves scientific productivity. Several studies have found correlations between use of CMC and papers published or similar measures of productivity. In addition, studies find that respondents attribute their productivity increases to the efficiency benefits provided by CMC. However, there has been little longitudinal analysis that would help untangle the direction of causality. The related, and more fundamental, question of the impact of CMC on the speed with which science advances is still very much unanswered.

Neither good measures nor any empirical data on whether the spread of CMC has led to a reduction in the time required to make new scientific discoveries are available as of yet.

There are several areas where new research is still needed. In addition to the issues noted above, there are also the issues of the effects of the Internet across countries and across disciplines, and its effects, in turn, on global science. We can observe significant differences in CMC use across disciplines. How these differences may be affecting the advance of science in different fields needs further research. Equally important is the need for good empirical work in the uses and effects of the Internet across countries. While the Internet has the potential to overcome many of the barriers of resources and geography, it may also be found that benefiting from the Internet is contingent on access to local resources that are not uniformly available. It is still very much an open question whether the Internet is reducing the differences in scientific productivity and access to the benefits of science across countries or whether it is exacerbating those differences.

NOTES

1. The author would like to thank Michael Lynch for suggesting this aspect of CMC.
2. See http://dpls.dacc.wisc.edu/apdu/lis_chart.html.

REFERENCES

ABBOTT, T. (ed.) (1994),
Internet World's on Internet, Mecklermedia, Westport.

ABELS, E.G., P. LIEBSCHER and D.W. DENMAN (1996),
"Factors That Influence the use of Electronic Networks by Science And Engineering Faculty at Small Institutions. Part 1. Queries", *J. Amer. Soc. Inform. Sci.*, 47(2), pp. 146-158.

ALLEN, T.J. (1977),
Managing the Flow of Technology, MIT Press, Cambridge, MA.

ALLISON, P.D. and J.A. STEWART (1974),
"Productivity Differences Among Scientists", *Am. Socio. Rev.*, 39, pp. 596-606.

BARLEY, S.R. (1990),
"The Alignment of Technology and Structure Through Roles and Networks", *Admin. Sci. Q.*, 35, pp. 61-103.

BIERLY, E.W. (1988),
"The World Climate Program: Collaboration and Communication on a Global Scale", *Ann. Amer. Acad. Polit. Soc. Sci.*, 495, pp. 106-116.

BISHOP, A.P. (1994),
"The Role of Computer Networks in Aerospace Engineering", *Libr. Trends*, 42(4), pp. 694-729.

BRUCE, H. (1994),
"Internet Services and Academic Work: An Australian Perspective", *Internet Res.*, 4(2), pp. 24-34.

BULLEN, C. and J. BENNETT (1991),
"Groupware in Practice", in C. Dunlop and R. Kling (eds.). *Computerization and Controversy*, Academic, Boston, pp. 357-387.

CARLEY, K. and K. WENDT (1991),
"Electronic Mail and Scientific Communication: A Study of the Soar Extended Research Group", *Knowledge: Creation, Diffusion, Utilization*, 12, pp. 406-440.

COHEN, J.A. (1995),
"Computer Mediated Communication and Publication Productivity among Faculty in Association of Jesuit Colleges and Universities (AJCU) Institutions", Doctoral dissertation, State University of New York at Buffalo.

COLE, J.R. and S. COLE (1972),
"The Ortega Hypothesis", *Science* 178, pp. 368-37.

COLE, S. and J.R. COLE (1967),
" Scientific Output and Recognition", *Am. Socio. Rev.*, 32, pp. 377-390.

CONSTANT, D., L. SPROULL and S. KIESLER (1996),
"The Kindness of Strangers: On the Usefulness of Weak Ties for Technical Advice", *Organization Science*, 7(2), pp. 119-135.

DONGARRA, J.J. and E. GROSSE (1987),
"Distribution of Mathematical Software via Electronic Mail", *Commun. ACM*, 30(5), pp. 403-407.

DUBROVSKY, V.J., S. KIESLER and B. N. SETHNA (1991),
"The Equalization Phenomenon", *Human-Computer Interaction* 6, pp. 119-146.

ECHAVARRIA, T., W.B. MITCHELL and K.L. NEWSOME (1995),
"Encouraging Research Through Electronic Mentoring: A Case Study", *Coll. & Res. Lib.*, 56, pp. 352-361.

EVERHART, D. (1996),
"The Labyrinth: An Electronic Information Network for Medieval Studies", in T.M. Harrison and T. Stephen (eds.), *Computer Networking and Scholarly Communication in the Twenty-First-Century University*, State University of New York Press, Albany, pp. 191-200.

FELDMAN, M.S. (1987),
"Electronic Mail and Weak Ties in Organizations", *Office Tech. People,* 3, pp. 83-101.

FINHOLT, T.A. (1993),
Outsiders on the Inside: Sharing Information through a Computer Archive (unpublished dissertation), Department of Social and Decision Science, Carnegie-Mellon University.

FINHOLT, T., S. LEWIS and W. MOTT (1995),
"Distance Learning in the Upper Atmospheric Research Collaboratory [UARC]", North Central Sociological Association Annual Meeting, Pittsburgh, PA.

FINHOLT, T. and G. OLSON (1996),
From Laboratories to Collaboratories: A New Organizational Form for Scientific Collaboration, University of Michigan, Department of Psychology, Working Paper (copies available from finholt@umich.edu).

FINHOLT, T. and L. SPROULL (1990),
"Electronic Groups at Work", *Org. Sci.*, 1, pp. 41-64.

GARRISON, A., P. WALKER and L. TERHAAR (1997),
"ULS Asks Scholars: Do You Need Us?", *Coll. & Res. Lib. News*, 58, pp. 544-545.

GARVEY, W.D. (1979),
Communication: The Essence of Science, Pergamon, New York.

GARVEY, W.D. and B.C. GRIFFITH (1979),
"Scientific Communication as a Social System", in W.D. Garvey (ed.), *Communication: The Essence of Science*, Pergamon, New York, pp. 148-164.

GURBAXANI, V. (1990),
"Diffusion in Computing Networks: The Case of BITNET", *Commun. ACM*, 33(12), pp. 65-75.

HAILMAN, J.P. (1996),
"Computer Networking in Ornithology," in T.M. Harrison and T. Stephen (eds.), *Computer Networking and Scholarly Communication in the Twenty-First-Century University*, State University of New York Press, Albany, NY, pp. 167-175.

HARMON, A. (1999),
"The Rebel Code", *NY Times* (21 February).

HESSE, B.W., L.S. SPROULL, S. KIESLER and J. P. WALSH (1993),
"Returns to Science", *Commun. ACM,* 36(8), pp. 90-101.

HILTZ, S.R. (1988),
"Productivity Enhancement from Computer-mediated Communication: a Systems Contingency Approach", *Commun. ACM,* 31(12), pp. 1438-1454.

HILTZ, S.R. and K. JOHNSON (1989),
"Measuring Acceptance of Computer-mediated Communication Systems", *J. Amer. Soc. Inform. Sci.,* 40(6), pp. 386-397.

HUFF, C., L. SPROULL and S. KIESLER (1989),
"Computer Communication and Organizational Commitment: Tracing the Relationship in a City Government", *J. App. Soc. Psychol.,* 19, pp. 1371-1391.

HURD, J.M. and A.C. WELLER (1997),
"From Print to Electronic: The Adoption of Information Technology by Academic Chemists", *Science & Technology Libraries,* 16, pp. 147-170.

IBRAHIM, Y.M. (1997),
"Finland: An Unlikely Home Base for Universal Use of Technology", *NY Times,* p. 1 (20 January).

JACOBSON, T.L. (1994),
"The Electronic Publishing Revolution is Not 'Global'", *J. Amer. Soc. Inform. Sci,* 45(10), pp. 745-752.

KAMINER, N. and Y.M. BRAUNSTIEN (1998),
"Bibliometric Analysis of the Impact of Internet Use on Scholarly Productivity." *J. Amer. Soc. Inform. Sci.,* 49, pp. 720-730.

KLING, R. (1995),
"Electronic Publishing and Changes in the Systems of Scholarly Communication", *Connecting and Collaborating: Issues for the Sciences* (NSF Workshop), San Diego, CA.

KNORR CETINA, K. (forthcoming),
Epistemic Cultures: How Scientists Make Sense.

KOVACS, D., K. ROBINSON and J. DIXON (1995),
"Scholarly E-Conferences on the Academic Networks: How Library and Information Science Professionals use Them", *J. Amer. Soc. Inform. Sci,* 46(4), pp. 244-253.

KRAUT, R.E., C. EGIDO and J. GALEGHER (1990),
"Patterns of Contact and Communication in Scientific Research Collaboration", in J. Galegher, R.E. Kraut and C. Egido (eds.), *Intellectual Teamwork,* Hillsdale, Lawrence Earlbaum Associates, NJ, pp. 149-171.

LEDERBERG, L. and K. UNCAPHER (1989),
Towards a National Collaboratory: Report of an Invitational Workshop at the Rockefeller University, National Science Foundation, Directorate for Computer and Information Science, Washington, DC.

MALTZ, M., A. GORDON and W. FRIEDMAN (1991),
Mapping Crime in Its Community Setting, Springer-Verlag, New York.

McGUIRE, T., S. KIESLER and J. SIEGEL (1987),
"Group and Computer-mediated Discussion Effects in Risk Decision Making", *J. Pers. Soc. Psychol.,* 52, pp. 917-930.

MERTON, R.K. (1968),
"The Matthew Effect in Science", *Science* 159, pp. 56-63.

MERZ, M. (forthcoming),
“Formen der Internetnutzung in der Wissenschaft” (“Forms of Internet Use in Science”), in R. Werle and C. Lang (eds.), *Modell Internet? Entwicklungsperspektiven neuer Kommunikationsnetze*, Campus, Frankfurt/Maine.

NATIONAL SCIENCE BOARD (1993),
Science and Engineering Indicators-1993, GPO, Washington, DC.

NATIONAL SCIENCE BOARD (1998),
Science and Engineering Indicators-1993, GPO, Washington, DC.

ORLIKOWSKI, W.J. (1993),
“Learning From Notes: Organizational Issues in Groupware Implementation”, *Info. Soc.*, 9, pp. 237-250.

ORLIKOWSKI, W.J. and J. YATES (1994),
“Genre Repertoire: The Structuring of Communicative Practices in Organizations”, *Admin. Sci. Q.*, 39, pp. 541-574.

OSTBYE, T. and T. WELBY (1988),
“Electronic Communication with Colleagues Overseas”, *New Zeal. Med. J.*, 101, pp. 145-147.

PHILIP, G. (1995),
“Use of ‘Leading-Edge’ Information Systems by Academic Chemists in the UK. Part 1. The Results of a Preliminary Investigation”, *J. Inf. Sci.*, 21, pp. 187-199.

RAINWATER, L. and T.M. SMEEDING (1988),
“The Luxembourg Income Study: The Use of International Telecommunications in Comparative Social Research”, *Ann. Amer. Acad. Polit. Soc. Sci.*, 495, pp. 95-105.

RICE, R.E. (1980),
“Impact of Computer-mediated Organizational and Interpersonal Communication”, *Annu. Rev. Inf. Sci. Technol.*, 15, pp. 221-249.

RICE, R.E. (1994),
“Network Analysis and Computer-mediated Communication Systems”, in S. Wasserman and J. Galaskiewicz (eds.), *Advances in Social Network Analysis*, Sage, CA, pp. 167-203.

RUHLEDER, K. (1994),
“Rich and Lean Presentations of Information for Knowledge Workers”, *ACM Trans. Off. Inf. Syst.*, 12, pp. 208-230.

SANDERSON, D. (1996),
“Co-operative and Collaborative Mediated Research”, in T.M. Harrison and T. Stephen (eds.), *Computer Networking and Scholarly Communication in the Twenty-First-Century University*, State University of New York Press, Albany, pp. 95-114.

SCHAUDER, D. (1994),
“Electronic Publishing of Professional Articles: Attitudes of Academics and Implications for the Scholarly Communication Industry”, *J. Amer. Soc. Inform. Sci.*, 45(2), pp. 73-100.

SCHWARTZ, C.A. (1994),
“The Strength of Weak Ties in Electronic Development of the Scholarly Communication System”, *Coll. & Res. Lib.*, 55, pp. 529-540.

SIEGEL, J., V. DUBROVSKY, S. KIESLER and T. McGUIRE (1986),
“ Group Processes in Computer-mediated Communication”, *Organ. Behav. Hum. Decis. Process.*, 37, pp. 157-187.

SPROULL, L. and S. KIESLER (1986),
"Reducing Social Context Cues", *Manag. Sci.*, 32, pp. 1492-1512.

SPROULL, L. and S. KIESLER (1991),
Connections, MIT, Cambridge, MA.

SUDWEEKS, F. and S. RAFAELI (1996),
"How Do You Get a Hundred Strangers to Agree? Computer-mediated Communication and Collaboration", in T.M. Harrison and T. Stephen (eds.), *Computer Networking and Scholarly Communication in the Twenty-First-Century University*, State University of New York Press, Albany, pp. 115-136.

TAUBES, G. (1993),
"Publication by Electronic Mail Takes Physics by Storm", *Science*, 259, 26 February.

UZUMERI, M.V. and C.A. SNYDER (1996),
"Information Technology and Accelerated Science: The Case of the Pentium TM Flaw", *Calif. Manag. R.*, 38(2), pp. 44-63.

VAN ALSTYNE, M. and E. BRYNJOLFSSON (1996*a*),
"Communication Networks and the Rise of an Information Elite", MIT Sloan School of Management, http://web.mit.edu/marshall/www/home.html.

VAN ALSTYNE, M. and E. BRYNJOLFSSON (1996*b*),
"Could the Internet Balkanize Science?", *Science*, 274, pp. 1479-1480 (29 November).

WALSH, J.P. and T. BAYMA (1996*a*),
"Computer Networks and Scientific Work", *Soc. Stud. Sci.*, 26 (3), pp. 661-703.

WALSH, J.P. and T. BAYMA (1996*b*),
"The Virtual College: Computer-mediated Communication and Scientific Work", *Info. Soc.*, 12, pp. 343-363.

WALSH, J.P., S. KUCKER, N. MALONEY and S. GABBAY (forthcoming),
"Connecting Minds: CMC and Scientific Work", *J. Amer. Soc. Inform. Sci.*

WALTHER, J.B. (1996),
"Computer-mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction", *Comm. Res.*, 23, pp. 3-43.

WALTHER, J.B. (1997),
"Group and Interpersonal Effects in International Computer-mediated Collaboration", *Hum. Commun. Res.*, 23, pp. 342-369.

WEISBAND, S.P. (1992),
"Group Discussion and 1st Advocacy Effects in Computer-mediated and Face-to-face Decision-making Groups", *Organ. Behav. Hum. Decis. Process.*, 53, pp. 352-380.

WEISBAND, S.P., S.K. SCHNEIDER and T. CONNOLLY (1995),
"Computer-mediated Communication and Social Information: Status Salience and Status Differences", *Acad. Mangt. J.*, 38, pp. 1124-1151.

ELECTRONIC PUBLISHING ISSUES

Table of contents

This article was written by Karel Leeflang, Managing Director, Elsevier Science, Ireland, and Jim Gilmore, Director General, CAB INTERNATIONAL, United Kingdom.

OECD 1999

I. INTRODUCTION

Information and the processes and systems used to manage and disseminate it are of essential and intrinsic importance to science. The science process is based on the utilisation of known information, and in turn delivers new information to the various communities of the world. The overall infrastructure of science must therefore provide for the systems and tools that enable information to be acquired, generated and disseminated, and these must be effective and sustainable. Furthermore, the systems and tools must ensure that standards of quality are preserved and that intellectual property rights are protected.

Electronic publishing is, of course, not a new phenomenon: computers have been used as a production tool in the publishing industry since the 1960s, and the derivative products have been made available through online systems for almost 30 years. The major changes have come about through the increasing availability and power of computers, including importantly the personal computer, particularly in the last decade. On the production side, this has affected publishers by enabling them to reduce costs of publishing and to develop a wide range of new electronic products. The range and capability of transfer media have also expanded, providing new opportunities for delivery of information resources directly to the user's desktop. We have seen the transition from floppy disk to CD-ROM to Internet transfer, and prospectively to the powerful new medium of Digital Versatile Disk, or DVD.

These technologies have not only opened up the delivery channels for publishers, they have also encouraged the emergence of competitive small publishers and enabled users to maintain and control their own publishing processes. The impact of all of these trends and developments has been seen in all areas, and is of particular significance to science.

II. SCIENCE AND INFORMATION

The information transfer and dissemination systems of science have developed and evolved over many centuries. They continue to be based on a combination of "formal" and "informal" systems: informal systems include many kinds of direct and indirect communication among scientists, while formal systems are based on established publishing mechanisms which are economically sustainable and quality assured. The transition which is taking place in conjunction with the expansion of availability of information technology resources is one where informal systems (and the information they make available) can proliferate, and where the formal systems must adapt to a new platform of scientists' needs and expectations.

Before looking specifically at the changes brought about by information technology, and the implications of those changes, we should address the intrinsic importance of information to science, and the systems used both currently and historically for managing information as an input to the R&D process, and information as a product of research.

Regardless of the delivery mechanisms and systems used, there is an essential need for quality, consistency, reliability and sustainability of these information systems. There has also been a long evolution of information delivery systems and media, and an environment of stability has been achieved. This will have to be re-established in the new world of information technology if the necessary characteristics are to be retained.

As indicated earlier, there are two principal kinds of system. The first can be categorised as informal systems, involving a variety of communication methods among scientists. As is the case for R&D in general, this kind of communication process is principally funded by governments, and to a large extent the costs are not overtly quantified, being "hidden" in expenditure budgets

While they represent an important element of the information transfer process, informal systems are vulnerable to funding cuts and loss of continuity as scientists change. They are also often susceptible to loss or weakening of intellectual property rights, and they are oriented towards informal and non-refereed information.

The second category of information transfer system can be categorised as formal, comprising published works of many different kinds. These works are produced, used and sustained by a mix of the public and private sectors, and they require direct and substantial expenditure within research budgets. They are more likely to be sustainable, particularly when market-based, although they are also vulnerable to changes in government funding for science, and they are susceptible to economic environments. It may be pertinent to comment that many international publishers have been significantly affected by the economic crisis in South-East Asia.

Importantly, the established formal information delivery systems deliver verifiable quality through peer-review processes, and provide protection for intellectual property rights through a combination of publishing conventions and national legislation.

The major significance to this infrastructure of developments in information technology is that the informal systems have the potential to explode, possibly to the detriment of the formal systems. The weaknesses in the former will then be magnified and the problems of quality, sustainability, and the other important characteristics can be multiplied. The issue is to recognise this and manage it, so that true benefits are obtained from IT, and that a new, valid and sustainable infrastructure emerges.

III. ELECTRONIC PUBLISHING TRENDS

The key element is the general international availability of PCs and associated equipment. From a global perspective, this has been particularly significant in "evening out" the technology gap between the developed and developing worlds. The new technology provides for expansion of the capability to store, process and distribute information and data. It is potentially of huge benefit in its impact, but can lead to a proliferation of invalidated information and data and presents serious issues related to sustainability.

The second major factor is the emergence of the Internet as a ubiquitous information delivery and communication medium. It is also of tremendous potential benefit, in terms of powerful new processing capabilities, immediate international information accessibility, access to new resources and media, and advanced methods of information management and access. However, it also introduces much increased potential for the distribution of poor-quality information, the same problems of sustainability, the risk of dependence on overloaded networks, uneven international accessibility, and some serious questions relating to the underlying economics of its use.

We will address and illustrate the trends in and scope of modern electronic publishing by reference to four principal aspects:

– Secondary databases.

– Electronic delivery of primary publications.

– Factual databases and databanks.

– Linking and integrating systems.

Secondary databases

Secondary services, including abstracts and indexes, have provided the primary means of selective access to the publications of science for more than a hundred years. They include the major services such as Chemical Abstracts, INSPEC (Physics and Electronics), Excerpta Medica and MEDLINE (Medical Science), CAB ABSTRACTS (Agriculture and the applied life sciences) and Biological Abstracts, together with a wide range and large number of services in other areas of the natural and social sciences. These services have been used throughout this era in the "traditional" form of printed publications and lately in electronic form through online services, on CD-ROM and now through Internet delivery.

Scientists worldwide are accustomed to and dependent on these services to provide a selective means of access to the vast body of published science, and they continue to retain a position of importance in the information transfer process, even as primary publications become increasingly available in electronic form. The secondary services are provided by a mix of private, non-profit and commercial organisations, with well-established pricing systems and models. These began as subscription-based services based on printed journals, which in fact continue to be used, and have progressively transformed to a model of ad hoc use and therefore transactional pricing across online systems, including the Internet.

For the most part, these services are required to be at least financially self-sustaining and, in some cases, profit-making. They are therefore necessarily based on sound economic models and principles and succeed (or fail) on the degree to which they provide added-value services to users and are therefore sustainable. This tends to be a volatile sector, with the major databases having long-term stability and sustainability while smaller and more specialist services tend to emerge and ultimately fail through lack of use or loss of subsidy.

The key, and as yet unanswered, question for the future is whether these services will continue to maintain a role in the face of perceived competition from direct electronic delivery of the primary publications. Our view is that they will continue to be valuable and in fact necessary as a "finding tool" or as part of an intelligent access mechanism to the plethora and diversity of delivery mechanisms and media used for supplying primary information.

Electronic delivery of primary publications

Elsevier Science is the largest science publisher in the world, publishing more than 1 100 titles. Some of the titles are published almost daily, many others are weekly. Elsevier Science is part of Reed Elsevier Plc, one of the top-ten publishers in the world with a strong programme in science, professional and business information.

Reed Elsevier owns some large databases, such as Nexis Lexis in business and legal information, and EMBASE with bibliographic information in the Biomedical field. Although Reed Elsevier and its companies have always had a policy of focussing on content and letting other parties take care of the distribution, due to a perception that some market requirements are not being fulfilled by third-party distributors the firm has now become directly involved in the development of its own dedicated search engines. This applies strongly to Lexis and Butterworth for legal information and Science Direct for science information. The advantage of

Reed Elsevier having a large part of the business-to-business information market is that this enables the development of search engines for multidisciplinary science professional and business user communities.

At the same time Reed Elsevier works closely together with software developers, for example it has entered a five-year contractual strategic alliance with Microsoft and various library consortia, academic institutions and other content suppliers and brokers. At the end of the day, it is always the user who will decide on the appropriate vehicle to obtain the best, most comprehensive and timely information. The content provider is the partner most able to organise and make that information available.

With the rapid developments in IT and the technical and economical availability of electronic access of information to the scientist's workbench, publishers have begun supplying scientific information electronically. Using the same technology as for the secondary databases described above, primary information has been delivered electronically to the market for quite some time. Although a number of programs and products are available on the market from other publishing houses, library consortia and distributors, we will use the developments into electronic delivery of primary information of Elsevier Science to illustrate the general developments of the scientific, technical and medical (STM) information market.

Elsevier began to experiment with delivering the full text of scientific journals on line in the early 1980s, building a database of some 25 000 pages and offering access to some 30 libraries in Europe and the United States. This work was technically successful, but far too far ahead of its time. User tests showed an interest in accessing such materials but there was clearly no market and delivering at 300 baud over normal telephone lines was inefficient. The work was therefore shelved pending the further development of the infrastructure and the market.

The next experimental programme for full text electronic delivery started in 1990, when Elsevier Science and a group of renowned American universities joined together in a three-year intensive trial and error programme called Tulip (The University Licensing Program) to study the technical and economical aspects of delivering electronic versions of Elsevier Science titles to the digital library environment. The service was inaugurated in 1991 with the delivery of bibliographic information and abstracts to one of the leading universities in Holland, followed, in 1994, with full-text electronic files for all the Elsevier journals to which the library subscribed. This was a logical step, the infrastructure was in place and access to full text was regarded as an appropriate step to improve our end-user service. The Tulip experiment did provide a vast amount of information to all parties. Technically it was far more complicated at the supplier and receiver ends, and economically it required a fair amount of investment at both ends.

These experiments resulted in the launch of Elsevier Electronic Subscriptions (EES) in 1995 which, after a somewhat slow start, is now becoming increasingly popular. Worldwide academic, industrial, national and governmental libraries are subscribing to this licensing programme. EES is delivered in four databases, table of content file, limited SGML files for fielded searching and hyper linking, flat text files containing flat ASCII text of all pages and items in each issue, and item files containing full text and graphics of all items published in ES journals.

Since 1993 Elsevier Science has also launched standalone electronic journals on the market: Nuclear Physics Electronic, Immunology Today Online, Creme-Combis linked with the European Molecular Biology Laboratory (EMBL), New Astronomy, and many others.

In 1997, Elsevier Science launched Science Direct, a Web-based publishing platform which offers libraries and their end users desktop access to the remotely stored full text of all ES journals, as well as full text from other participating publishers. Linkages to document delivery services, commercial databases, reference works, newsletters, textbooks and gateways to other information types and sources are fully available. This product/program fully serves the needs of the research world – one-stop shopping, access to all relevant information for the specific need of the researcher, and problem solving – therefore improving the efficiency, and consequently the economical situation, of the research world.

Elsevier Science is not the only science publisher offering these tools for electronic access and retrieval of information to the science world. Other similar programs/products are available on the market, some with the same but others with less sophistication. A number of these experiments have not survived and, in the absence of sufficient critical mass and funding, there is no way to serve the market on a continuous basis. In addition, as the user communities also have to invest heavily in infrastructure, hardware and software, it is vital for them to take up the right partnership with their content suppliers. Although commercial publishers are not always popular with library institutions, it is fair to say that the initiatives and developments into electronic delivery of primary information coming from the private sector do meet the needs of the user communities. It is clearly difficult for not-for-profit and not-subsidised content providers to afford these huge investments into sophisticated computer-aided production of manuscripts including conversion to SGML or HTML structured files, and linkages to other information sources. They also lack the critical mass to justify these investments, and the revenue on electronic products is still very small.

Most of the forecasts, for example from the European Union, are very optimistic but the reality is that electronic revenue for most of the traditional STM publishers represents less than 15% of their total revenue. It is growing but is in no way sufficient to finance the necessary investments.

There is also the change in economic models. Libraries are used to paying for subscriptions and are responsible for making the information available to their end users. They have built their present budgeting and funding systems on this pre-paid strategy, as have the publishers, but the delivery of electronic full-text files, directly or remotely, to the end user's workbench will mean that new systems of payment and budgeting will have to be developed.

However, most publishers are currently offering content to their institutional customers on a licence agreement basis. Linked to the subscriptions of printed journals, electronic files are delivered at a premium fee with certain usage conditions. In this way, the library is both able to budget and can become more cost efficient in the delivery of journals and documents within the institutional environment. Single electronic files not linked to a printed journal subscription can be included in the licence.

In future, both parties, libraries and publishers, will continue to work on a subscription basis for the availability of information linked to the use of individual articles. Using sophisticated software, the user will have access to the vertical information, the journal and the horizontal information in his field of research.

Factual databases and databanks

The advent of the Internet and sophisticated electronic information systems has stimulated the growth of a new range of databases. Ten years ago, if one referred to a scientific database, then it was probably a secondary or abstract database which allowed the user to scan through thousands of journal abstracts to find the primary article required. However, in areas such as biotechnology, there has been an almost unparalleled growth of "factual databases" such as the EMBL Nucleotide Sequence Database which stores annotated nucleotide sequences (the DNA/ genetic sequence database).

Databases are essential tools in modern research. They are manipulated and analysed, and data is taken from them to have value added to it. Thus, we already have "databases built from databases" which "translate the genetic sequences into amino acid sequences", and others which enable the user to fold these amino acid sequences into protein models. We are now beginning to translate the raw sequence information into functional proteins.

Such databases – small and large collections of data – are usually assembled by scientists to fulfil their own needs and those of their fellow scientists. Hundreds of databases are being built, mostly in the public domain, and the best and most used are offered to the public via centres such as the EMBL Outstation, The European Bioinformatics Institute in Hinxton, Cambridge, or the many European Molecular Biology Network nodes (EMBnet) spread throughout Europe.

Informal figures from these centres show that these novel databanks take at least three to five years to reach sufficient critical mass, and require three to four full-time people working on the collection and validating the information and distribution mechanisms to make the information accessible and useful for the science world. However, up to 50% of these databanks disappear before they have been in existence for five years. Those which survive are presently continuously subsidised by national or supranational bodies. A very few (around 1%) look to private resources for their future.

Such a rate of closure is disquieting; at the least it means that money and effort were wasted on projects that had no future, but it can also mean that important data is no longer available because the database was established in an insecure setting. Solving this problem is obviously an important task for the public funding authorities but the growth of this sector raises concerns. Is there enough public money to maintain these rapidly expanding databases? And what funding can new databases expect? It seems clear that an improved information and co-ordination system surrounding this field is required, detailing what databanks are available in the various scientific fields. Ownership, funding, access to the data and overlap between databases are all issues that could be better examined and reported.

The fear of most such databases is that they will lose their public funding or that they will be unable to keep abreast of the new technologies needed to remain competitive. At present, in the United States, public money is available for such databases as MEDLINE, AGRICOLA, GENBANK and a host of biotechnology databases. The EMBL, and to a great degree, the EU, fund the EBI databases but SWISS-PROT, one of the most central databases in the biotechnology field, is now having to find alternative ways of raising income as it can no longer be totally supported from public funds. If this is happening to such an icon database, it will soon be happening to many others.

The question is, can they be privatised? Not really, unless the user is willing to pay for valuable information and provided there is no competition from similar fully subsidised databanks in other parts of the world. Science is, by definition, a global activity but access to scientific information is of vital importance to each national or supranational entity. Dependency on restricted accessibility to databanks or non-sustainable databanks is a serious threat to the development of research. Our hosts of today have an important role to play in the modelling of existing databanks, and at the same time creating an environment in which public and private databanks have an economic existence and continuity.

Database subsidies

Many databases receive government subsidies, especially in the United States where some products have become synonymous with “free information”. This clearly distorts the marketplace but also affects the scientific environment which is dependent upon accurate and broad-based information.

Although competing with subsidised products is always difficult, the situation was more bearable before the Internet effectively removed geographical borders. Today, products which were previously considered "national" are distributed or accessed all over the world, seriously distorting other national markets.

An excellent example of this phenomenon is MEDLINE – a well-respected product which can be used more or less for free the world over. As such, it not only challenges the commercial database products that compete with it, but also the not-for-profit databases which similarly serve this market. Furthermore, its availability on the World Wide Web means that many database hosts and distributors face a loss of income and use as users tend to move to the government-financed database hosts and ignore European (and American) commercial database hosts which have previously offered this and other databases to their customers.

While some users praise the actions of the US Government, others realise that one cannot rely upon any one database for information research needs. MEDLINE provides indexed abstracts of some 3 000 or more journals. This is only a moderate percentage of the total number of journals relevant to biomedical research. Furthermore, the journals covered in this selection are quite naturally more likely to be American titles, and so users are therefore primarily taken to American primary journal sources while other European and international titles are ignored. The support of the US Government also confers a degree of authority which leads to the perception that the best journals are covered in MEDLINE; and therefore that the best journals are American, irrespective of the truth.

Equally important is the fact that, while any search on MEDLINE might locate one or more "hits", it will almost certainly miss many others covered by competing services (for instance, the overlap between MEDLINE and EMBASE – probably its foremost commercial competitor – can be as little as 40% in some subject areas). The false sense of security such a result offers can be scientifically misleading as well as expensive, as many researchers might follow lines of research which have already been covered by others and published in equally respectable journals which were not abstracted by MEDLINE. (It is significant that researchers who have to be sure they have covered all the relevant literature, for instance those in pharmaceutical companies, always use a series of databases in addition to the subsidised American products.)*

Furthermore, because MEDLINE is cheap and freely available, it is increasingly used by many other more factual databases as a "linking tool" (*i.e.* they refer from

---

* For an independent evaluation of this issue see, for instance, Cecilia, M. Brown (1998), "The Benefits of Searching EMBASE versus MEDLINE for Pharmaceutical Information", *Online & CD-ROM Review*, Vol. 22, No. 1.

their data to MEDLINE). This increases MEDLINE's own visibility but also greatly favours other American-funded databases that are also linked to MEDLINE files. Basically, the user is taken to American secondary databases and journals rather than being left to look for the best product in the area.

Overall, the user might actually suffer through the free distribution of this database; and the US Government also subsidises other databases in other specialities such as agriculture, so this pattern is being repeated in other scientific disciplines. International science would suffer enormously if the complementary databases that compete to some degree with these subsidised products were to go out of business. Then the user would only have access to a small part of the accepted literature needed for good R&D, a situation that would benefit no-one. The OECD is one forum where this complicated issue could be discussed to ensure that a better balance is obtained.

Linking and integrating systems: the ADLIB project

For the future, the real power and potential of the use of modern information technology will be seen in the way the systems, services and databases described above can be linked and integrated to provide even more useful and usable information access methods for scientists. The combination of the Internet, powerful PCs and modern software tools can enable the scientist, from his desktop, to access a wide range of information and data resources, in ways which were simply not possible in the past. These new kinds of services will require new economic and pricing models and, importantly, greater collaboration and partnerships among information suppliers. Some of these services are already beginning to emerge, and the following illustrates one of the approaches which has been taken.

The Advanced Database Linkages in Biotechnology (or ADLIB) project is a joint venture of some 12 European publishers and data and service providers in the European bioinformatics sector. It has been supported by the European Union as a demonstration project, designed to show how the new technology can be used to improve the availability to scientists of European information and data relating to biotechnology.

The concept is based on the innovative use of technology to link various and diverse databases across the Internet through a seamless interface. At the simplest level, for example, a scientist can interrogate one or more bibliographic databases through an intelligent interface, identify references to a gene sequence, and subsequently link automatically to the database which contains the full sequence data. We have, therefore, for the first time, a tool that can integrate a substantial body of references into the published literature and allow this to be linked to comprehensive and detailed factual data, in this case on gene sequences.

The power and potential of this approach is evident: it suggests that scientists across the world can be provided with access to information retrieval and analysis tools which are far in advance of anything that was available in the past, and the concept is transferable to virtually any area of science.

IV. MAXIMISING THE BENEFITS: OPPORTUNITIES AND CHALLENGES

In the publishing context, the challenge is to use the technology and information resources to best effect in order to support the R&D process effectively, to improve the quality and cost-effectiveness of science, and to improve dissemination of results.

This must necessarily be in the further context of the total R&D infrastructure. The role of governments (in particular) is to ensure that all of the support systems are sound, viable and sustainable. Information technology impinges on all of the formal and informal systems for acquiring, using, generating and disseminating information.

The change presents undoubted opportunities, but brings with it threats to the established order and to this element of the fundamental infrastructure of science. It is an evolution which can neither be left to find its own level and mechanisms, nor to develop as an unchecked free market phenomenon. It raises important issues for government policy, research management considerations, questions for the academic community on the fundamental principles of the research processes, and, of course, many economic considerations.

The opportunities, and hence the benefits, of the use of the new information technology are clear and will be seen in better and faster information delivery, enhanced international co-operation, the potential for rationalisation of programmes based on greater access to information about other international research, and, ultimately, more rapid scientific progress.

Some of the key aspects can be seen in the trends and developments described above, where new and more powerful services and delivery mechanisms are emerging, and where in particular there is a potential for much greater integration of services and increased collaboration among the service providers. The combination of the new services and resources and strategic use of the technology will enable scientists and science managers to develop research programmes on a more dynamic information and communication base, with greater awareness of the importance (both scientific and economic) of the information systems which are applied to the science process.

To a certain extent, these new services will reduce reliance on the maintenance of collections of printed material, with the attendant cost implications, but they will at the same time introduce the need for new disciplines in the management (and cost management) of the use of electronic services. They will allow better and more rapid dissemination of research results around the world, and will facilitate communication and interaction among scientists. Importantly, they will make accessible information that was previously inaccessible, or at best accessible only at high cost and with difficulty. As noted above, the new era will bring increased integration and linking of data resources, allowing more powerful data retrieval, analysis and modelling.

Among the many concerns generated by this massive increase in the scope of electronic publishing are: the methods and mechanisms for retaining and assuring the quality of published science, which to date have been based on a very effective peer-review process; the management and security of intellectual property rights which, again, are subject to formal and well-established publishing mechanisms; and the crucial function of archiving the results of scientific research, where there is a real danger of a virtual breakdown in the system because electronic information proliferates more rapidly than the timely availability of the required archiving systems.

The provision of information on the Web is a great joy to many and a great concern for others. Researchers in the science community can easily talk to each other about research programmes and their latest findings. Many like to use the Internet to “chat” about new developments and there are some who feel that these “bulletin boards” will eventually replace the journal. There is no doubt that bulletin boards have a function, but scientists are equally insistent that science has to be validated.

The Ginsparg pre-publication service is sometimes used as an example of how Internet services can succeed. However, two points should be taken into account: first, Ginsparg “piggy backed” the refereeing system of the primary journal – only placing a pre-print on line when the paper had been accepted by a recognised journal; second, the service was free. I will not go into the ethics or otherwise of using the journal refereeing system for one’s own service but clearly someone was paying for the refereeing which is seen as essential.

Scientific information services have to offer quality, continuity and reliability. Increasingly, they have to be fast and efficient. As long as the academic researcher is not worried about the time taken to find and select relevant and good-quality information, he or she will use the free flow of information. However, there is no doubt that, as time pressures increase, the academic, like the industrial scientist, will want to rely upon pre-selected, validated information to shorten research time and to improve cost efficiency.

Publishers provide this service and will use the Internet to facilitate the distribution of useful scientific information. They will use this as an extension of what they are doing now – bringing peer-reviewed materials to the user. Journals and literature databases will increasingly be integrated with the factual databases mentioned above so that users can benefit from a linked system of information resources offering quality materials.

Yet another trend in the need for validation should perhaps also be mentioned. Increasingly, the peer reviewer is being asked to judge papers which rely upon the analysis of data that can only be done by computer. To date, there is no obligation for scientists to deposit such data as they write their paper and the reviewer has to go some way on trust. In the future, it is clear that "supporting" or "subsidiary information databases" which can be used by the research community to double check original decisions and measurements will need to be established. This system already exists in the nucleotide sequence field and will soon be extended to other areas. Again, funding will be an issue, although the EU is interested in examining ways in which public and private funding can be used to handle such a combination of tasks.

Content suppliers, databank owners, governments and research communities are constantly discussing the issue of intellectual property issues in a digital environment. The words *intellect* and *property* do not seem to belong together. Intellect is personal and private. Property concerns boundaries and interaction with the world. So intellectual property has to do with a right to recognise the individual author/creator as well as the right to prevent trespass.

In the STM world, it is well accepted that copyright is transferred from author to publisher. Authors care about being distributed, they care about the ideas represented in their papers, the use in their institutions of their papers and the sharing of results with their collaborators – none of which conflicts with the commercial interest of the STM publisher.

STM publishers are keen to protect the integrity of the work of STM authors and so the intellectual, as well as the property, rights are seen as being complementary. It is not so much the author-publisher relationship which causes food for heated discussion, but more the boundaries to be set for fair use, and the need to protect the publisher's economic rights.

Thus, the publisher needs the neighbouring right, to protect his investment in organising and laying out the work (already protected under UK law), or he must be able to contract or licence the right to be able to sue for infringements of his economic or industrial rights. This should also cover database protection in order to protect the skill and investment put into organising a database. Currently, most discussions on copyright focus on the "creativity" aspects rather than on the value in terms of usefulness and comprehensiveness.

In Europe the *sui generis* right (meaning simply *new* or *different*) of the EU Database Directive is one approach – and a creative one – to give database producers an incentive to produce more databases in Europe. The Directive notes that the protection is for the maker of the database in order to protect its investment.

In contrast to the EU Directive, the US database protection which has passed the US House of Representatives and is now in consideration before the US Congress, is based on misappropriation (unfair competition principles) rather on than the creation of a new industrial right. The European Union and US approaches, however, have the same result. Both protect the proprietor against unauthorised misappropriation.

Publishers and their associations are very active in defending their intellectual property rights, at the same time preparing and proposing neighbouring rights to protect their business, their investment and their continuity. Of course, public opinion sees this as being purely commercial, but there is no difference between commercial and non-commercial publishers, both see it as their task to enhance the flow of information from scientist to scientist on an economic, quality-secure and continuous basis.

However, publishers do not do their business, and especially their electronic business, predominantly on a legal platform. The experiments mentioned above have provided publishers and user communities with a great deal of information about technical and commercial issues. Based on this fact finding, licence programmes have been set up between the two parties covering all aspects of delivery and usage conditions. In this way, both parties know what to expect, and no party has to fear that there is no control on the use of information in those communities.

Scientific information is *not* at a cross roads, in spite of what some critics say. It is facing some exciting challenges but it should be able to rise to meet them. Increasingly we will see the need for public and private funding initiatives to interact. Basic research discovers basic facts, which are used by scientists for applied and/or further intellectual work. The present premise is that this basic, raw data should be in the public domain (*e.g.* the EMBL Data Library), but the value that is added to such data can surely be charged for. Adding value means investment, and investments should be protected.

There is no doubt that we need an environment in which subsidised projects are not able to compete with unsubsidised ones, and where databases can evolve from the base R&D level at which they start life into exploitable products (whether funded from the public purse or by the user subscribing to them). The protein scientist looking for, say, new molecular structures in the search for a drug, might today look at some 30 databases from a series of 100 available across EMBnet. He/she will check the data in the secondary databases, and review other findings in the original journals. All this must be done in a verified, stable, sustainable and trusted environment. The publisher has a key role here and will continue to play it.

This is only part of the solution, however, because we will continue to make electronic content easy for users to find, access and trade, while at the same time easy for publishers and libraries to trace and account for. We call that trading electronic content. The answer has to be through standards. Standards must be wider than any one community. For example, publishers are anxious to retain their brand identities. But this must be done at a higher level than simply a proprietary means of communication or interface. We need standards in the basic enabling technologies to give us a secure basis on which to be creative and add value. We need standards in a number of areas: we can wait for them to happen but publishers have chosen to be involved because they are essential for the future of our industries, including all players in the chain. These standards should cover: identification, descriptive data, rights data, presentation and exchange formats, user identification, and authorisation and rights clearance.

Much work remains to be done, but there is a lot of co-operation and good will. One example of this collaborative approach is the development of standard serial document identifiers in the Digital Object Identification (DOI). At the STM meeting in Frankfurt in October 1997, this project was accepted by the entire STM publisher community as well as by the photographic and music industries. This is especially important because all information in whatever field and format could potentially be treated as a digital object. Perhaps fortunately, from a business point of view, we do not yet have a homogeneous digital publishing world; there are still very definite differences between text images and sound, for example. But the underlying technologies and standards must converge: we want, and our users need, to put images of chemical structures up with chemistry text and perhaps phonetic clips with our language encyclopaedias. Technology must allow that to happen but in practice that means standards are needed to be able to make full use of the Internet and its likely future development.

So, trading electronic content requires an infrastructure that we can all use with confidence, so that we can use our energy to build creative new solutions and add value to our products, rather than spending unnecessary time converting file formats and overcoming barriers. Setting those standards will not be a trivial effort: it will require commitment and time, as well as participation in discussions like this, in committees if necessary and in pilot projects and prototypes. We will get out of the process only as much as we put in.

This represents a tremendous challenge and opportunity. By initiating the project and collaborating for the mutual benefit of the scientific information industry and by spending time defining effective standards, we can ensure that the infrastructure we build will be one that the whole scientific world can make use of for the benefit of science.

V. CONCLUSION

We tend to see science in terms of its intellectual processes and its contribution to the development of new techniques and technologies. To some extent, we may overlook the vast networks and systems for information transfer which underpin science, and which represent a very substantial cost.

There is a long-established order of scientific communication, based largely on traditional and conventional publishing processes which are now being challenged, to adapt to the impetus of electronic publishing, as a result of the dramatic expansion in the power and availability of information technology. This article has illustrated the trends in the publishing industry, and described some of the strategic and economic issues that currently confront the science community and its supporting publishing infrastructure.

The opportunities are undoubtedly vast and far-reaching, and the ultimate outcome will be a new environment which will be strongly conducive to assisting and enhancing the research process. It will, however, require some reassessment of the roles of publishers and in fact the roles of scientists themselves in the information transfer chain. It will also lead to some fundamental reconsideration of the costs and economics of scientific communication, with increasingly important impacts on research funding.

The various issues raised here, and in particular, aspects such as standards, intellectual property rights and the economics and sustainability of electronic publishing services, will become increasingly relevant and important to research managers and funders, and no longer principally the province of publishers. It will be necessary for all of the communities of science to work together in order to develop a viable and sustainable platform for this new era of electronic publishing.

DENMARK'S ELECTRONIC RESEARCH LIBRARY: A TOOL FOR INSTITUTIONAL CHANGE

Table of contents

This article was written by Jens Thorhauge of the Danish National Library Authority.

I. INTRODUCTION

Rapid access to information – the raw material for knowledge – is a crucial strategic factor in many fields, particularly research, education, business. The World Bank recently stated that knowledge is the single most important factor for welfare states. The Information Society is built on the intensive production and use of information, and is characterised by a growing number of citizens who in their daily work require access to professional information.

The importance of electronic information stems from the fact that, via networks, more people can access more information more rapidly. And, if libraries combine their traditional services with net-based services, the potential benefits for knowledge are even greater. This is, of course, highly relevant for researchers, but the impact on the general upskilling of the work force and on public access to and use of information will also be significant. Highly developed networked research information systems are crucial to the fostering of a productive global research village.

Libraries worldwide are undergoing dramatic changes, becoming increasingly digitised and moving away from their traditional paper base. In European and American research libraries in particular, tremendous efforts are being made to implement new IT-based services.

Over the last two decades, this information revolution has taken place at growing speed. The first stage was dedicated to automation: with IT being used for housekeeping functions such as circulation, acquisition, cataloguing and stock control. The next step included online access to databases, typically through union catalogues which constitute a single point of contact for multiple library holdings. As net-based services developed, libraries started to use the Internet. The final step towards the digital library includes Internet services via the library's homepage for remote use, access to full-text databases, Internet guides, ordering facilities, electronic document delivery, targeted services in the fields of research, business, education, IT training. This development affects the organisation and the professional roles, services and goals of libraries. The basic change is reflected in the movement away from a collection orientation to a connection and user orientation. Behind this statement lies a profound change in professional culture and values.

The vision of the digital library at the beginning of the 1990s was straightforward: digitised collections provide the opportunity to access the library’s resources from any computer with a connection. The ultimate objective is direct access to information or information at the users’ fingertips. As the Internet has expanded, different terminology has been applied: electronic library, virtual library, digital library – with the latter probably the most precise. Recent efforts have clearly shown that the idea of purely virtual libraries, *i.e.* a single completely digital system, is unrealistic. What we are aiming for today is quicker and better access to more research material for a larger number of researchers and students.

In recent articles, digital libraries are perceived as a collection of disparate resources and disparate systems, catering to specific communities and user groups, and created for specific purposes. Especially within the humanities and social sciences, they will probably indefinitely hold large paper-based collections.

Digital libraries face some key challenges: building an adequate technical architecture; providing the critical mass of digitised content that will attract users; establishing standards; dealing with copyright management, which in the digital world is more difficult to handle because copies cannot be controlled. The preservation of digital material is a major unresolved issue. In conclusion, building digital libraries is difficult and expensive and demands a long-term and sustained effort. The challenge at the national level is to build a co-ordinated digital library scheme in which established libraries co-operate and provide networked access to their disparate systems and resources through a single interface. All modern libraries provide access to digitised material; they now have the opportunity to enhance this service by entering a new type of co-operation (which may be painful as it involves relinquishing part of their right to self-determination). Technological change inevitably leads to organisational change due to the fact that an increasing part of the research libraries’ collections must be shared for economic reasons, and because users will insist on it, as technology offers brilliant opportunities for just that concept. Predictably, the independence of the institutions will be reduced, but it is not yet clear to what degree.

The shorter-term perspective is rapid and easy access to articles and bibliographic information for a broader public. Consequently, the potential exists for a qualitative leap in library service that should allow them to more efficiently reach their targets – to the benefit of all library users.

Another important aspect is the consequences for institutional relationships. Co-operation on digital resources should evolve into co-operation in other fields, for instance housekeeping systems, and should lead to rationalisation.

This article focuses on Denmark's Electronic Research Library as an example of a strategy for changing the relationships between institutions, developing new services and turning a project into a permanent service without project funding.

II. DENMARK'S ELECTRONIC RESEARCH LIBRARY

Danish research libraries will be developed over five years (1998-2002) to function as a single, integrated digital research library: Denmark’s Electronic Research Library. In co-operation with the Ministry of Research and the Ministry of Education, the Ministry of Culture will be investing Dkr 200 million in this project which aims to provide researchers, students, business, and other professionals with easier, faster and more effective access to the latest research information. The project is part of the government’s new initiative for research and IT, and in the 1998 budget the project has been allocated Dkr 126 million. Furthermore, the three ministries involved have allocated a further Dkr 74 million, bringing the total to Dkr 200 million.

The concept

The network of research libraries will form a virtual system which will transcend the borders of regional and local libraries in a simple, transparent way, and within the given legal and economic framework. It will make the libraries’ collective information resources of digital and traditional materials available to users all over the country.

The establishment of Denmark's Electronic Research library is based on existing technology. The aim is to provide an effective national information supply, which makes available resources already developed in research libraries throughout the country, and enables other information centres to be incorporated into the virtual information system. The Electronic Research Library will emerge as a single, large, coherent, virtual information system. The overall effect is achieved by linking the networks of the research libraries and by complying with standards for communication, for search support, for registration and indexing, for document description and representation, etc.

Electronic access to catalogue information and to an increasing number of digital documents will be among the new features presented to users. More specifically the project will provide:

- Rapid and co-ordinated searching across several databases, based on a combination of numerous search profiles, independent of the physical location of the databases, both nationally and internationally.
- A uniform, user-friendly retrieval system with the same user interface, whichever database one is searching in.
- An automatically updated list of literature and periodicals, displayed according to individual search profiles and criteria using robots.

The system will be accessible by all: from the home, from the office or from the library.

The content of the Electronic Research Library has been thoroughly discussed in a working party report delivered in February 1997. The first concern of the report is to establish a "critical mass" of library materials. This requires making available a vast amount of resources in such a way that they are easily and rapidly accessible to researchers. Once the benefits become obvious, they will begin to use the facilities. The report identifies three types of material that are relevant in establishing the Electronic Research Library, and to which high priority should be given in the project phase:

- Full-text databases, mainly journals, made accessible through licences with the large publishers.
- Reference databases.
- National bibliographies, primarily the Danish DanBib-base.

High priority should also be given to tools that will link the virtual library to the physical library, *i.e.* retro conversion of catalogues. If the appropriate tool is chosen, collections of physical items can be partially digitised. Obviously, the first step towards establishing a "critical mass" is to provide access to the greatest number possible of relevant journals.

Three factors can be used to describe the concept of Denmark's Electronic Research Library: the challenge, the objective and the means.

- The *challenge* is to cope with the increase in the production of research information and with the essential role that the use and the application of this information plays in the development of society, education and business.
- The *objective* is clear: Denmark will achieve the qualitative and economic advantages of digital – and network-based research libraries – offering users access to relevant research information directly, regardless of where the information is located.
- The *means*: A network of research libraries and information centres linked together in a coherent structure providing researchers and users with research results more easily and more rapidly, and offering more effective access to the most up-to-date research information.

The Danish vision includes a platform, where not only researchers at the twelve largest research libraries will be able to access an extended mass of electronic journals, but maximum benefit from national licences will be offered to smaller research libraries and libraries at minor research and educational institutions. In a further development, county libraries will have access on a "pay per view" basis – and, in due course, all libraries in Denmark should be connected to the Electronic Research Library.

The vision includes new services based on individual user profiles and the development of intelligent agents. A special benefit might be that the Electronic Research Library could be developed to make available not only a full catalogue of

Danish research, but also a full-text representation of all Danish research. This idea is closely linked to the publishing policy of universities and researchers in relation to electronic and institutional publishing codex. The potential exists to establish an alternative publishing structure – but this is in a longer-term perspective.

To turn this concept into reality, the Danish project has identified four main components – each of which is necessary to achieve the objective:

- The national infrastructure.
- The library infrastructure.
- Digital resources.
- User facilities.

The national infrastructure

The national infrastructure is the IT network and facilities enabling libraries and users to communicate efficiently. The Danish Research Net has been chosen as the IT network. This IT network has an advantage in that a substantial part of the users are already connected. This national infrastructure requires, however, more than just technology. The overall infrastructure requires adopting common guidelines for, in particular, national licence agreements, exchange of information, use of international standards, unified user access, etc. The regulations regarding user administration should also be coherent and consensual.

Library infrastructure

To enable individual libraries to become components of this joint, virtual library requires a certain amount of modernisation. Until now, it has been acceptable for each library to use its own IT systems and organisational procedures. The establishment of a virtual research library requires the standardisation of the technology and a number of organisational issues.

Implementation of the concept will require new technology. In order to obtain maximum benefit, libraries must change their habits and base their co-operation on common standards and guidelines.

Denmark's Electronic Research Library must create a profile and a level of service which encourages users to feel that they are "customers" of a common virtual unity, rather than of physically separate research libraries with different conditions and cultures. A similar profile should be presented to suppliers. It is, therefore, necessary to develop a common set of rules to formalise this co-operation. Examples include:

- The move to direct ordering from any library will change library co-operation procedures and the economic foundation linking institutions and appropriate ministries.

– The acquisition of a single digital copy will require co-ordination of, for example, agreements on national licences.

The inclusion of the whole body of research library users will necessitate that the level of service is determined according to common guidelines.

Increased co-operation among research libraries will require global management and co-ordination. Co-operation must be established across ministerial borders, although local participants need to retain their independence in order to preserve the dynamics of the system.

Digital resources

For traditional "non-digital" materials, the most important issue is the catalogue, although digitisation of some parts of collections is under consideration. For "new" materials, which are delivered in digital form, the challenges include: efficient management; widespread access; protection against damage and misuse; and migration to future technological platforms.

The document servers should initially provide the necessary storage capacity for text-oriented documents in HTML or PDF format. Later, storage capacity for larger amounts of scanned and multimedia documents will be required.

User facilities

Facilities will be a major issue for digital library users – notably, an economic issue. It will be crucial to provide users with sufficient facilities: workstations, access to printers, and software for information retrieval and manipulation.

In this context, both Personal Computers and Net Computers are under consideration. The choice is difficult to make at the present time. Personal Computers are familiar to libraries and the majority of users, and all types of software are available for standard Personal Computers. On the other hand, Net Computers may be an attractive option – in particular if the promises relating to security, lower cost and, especially, less time spent on system administration are fulfilled.

Organisation and plans

The investigation

The project was defined in a project description of September 1996 by the three ministries involved. An *ex ante* evaluation of the project was carried out by UNI*C and Ernst and Young. The report was delivered in Spring 1997 and an English-language summary is available.

The evaluation focused on the information technological aspects and deals only briefly with the development of the processes and organisational consequences. It singles out the factors that will change processes and organisation, but does not present any recommendations or advice as to how the changes should be implemented.

Organisation of the DEF project

The DEF project will initially consist of the twelve major research libraries, although in principle all libraries will be invited to participate.

The project to manage the implementation of Denmark's Electronic Research Library is organised as follows:

- A *co-ordinating board* consisting of members from the three ministries.
- A *steering group* with ten members appointed by the ministries and organisations.
- A *secretariat* that is integrated in the Danish National Library Authority.

The steering group represents various competencies: library management, research and IT.

The role of the Danish National Library Authority will be to execute the decisions of the steering group in general; and, in the licensing area, it will be the formal holder of national licences.

A national licence will be defined, not in respect of users, but in relation to state economic support. According to the report's proposals, any library can co-operate with another library or institution, form a consortium and negotiate licences. The consortium can apply for support from the Danish National Library Authority; this body will ensure that all relevant institutions are invited to share the licence.

In connection with the organisation of the Electronic Research Library, a task force will be set up to deal with national licensing, which for the moment seems to be the most important model for rights management.

Pilot projects

Six pilot projects have been initiated for the period 1997-98, with a budget of Dkr 8.2 million. These projects cover: communication, digitisation, digital archiving, and a wide range of users from libraries throughout the country.

The six projects are:

- A virtual faculty library covering all digital information on the social sciences. (The State and University Library Aarhus, Dkr 2 million.)
- Making the whole CD-ROM collection of the Royal Library accessible for a wide range of users in public libraries and university libraries with a fast response time.

- Establishing a test installation of digital research information on a large scale for a wide range of users on the local network.
- Digitising research literature on the subject of communication in order to establish a digital archive for research and education.
- Establishing a fast and continuing information system for electronic publishing of historical research publications as a co-operative project involving a university, a university library and a university publisher.
- Building up and promoting a reference information database of research articles combined with full-text research literature.

Status of the project at the beginning of 1999

- The national infrastructure is up and working.
- A plan for upgrading the library systems of the twelve largest research libraries has been established and the process has started. A further plan for the upgrading of 44 research libraries has been approved and the process has started.
- Digital resources are growing rapidly. National licensing has provided access to journals in practically every field.

Digitising older catalogues is a huge task – some have been done, but an overall plan is close at hand.

In the same way, a plan for the digitisation of existing collections has not yet been worked out; this task is considered in the long-term perspective. Since the digitisation of collections is expensive, tools have to been found which enable the appropriate components of documents to be selected, *i.e.* the parts in which users have a special interest and for which digitisation presents obvious benefits.

The most demanding challenges are linked to designing an efficient user administration, to designing a logical and self-explanatory interface, and to rights management. In this context, users should identify themselves in order to obtain access to licensed material, users should be easily guided to the subjects they are searching, and rights management should not be an administratively unbearable burden.

The steering group is currently discussing the future organisation of the research libraries. The strategy underlying the five-year project is to accelerate the inevitable change of paradigm. The extra funding is obviously useful in upgrading systems and providing access to electronic journals. However, if the project stops there, all that will happen after the five years is a demand for a new project to support a never-ending need for upgrading. What is required is a system that is able to cope with another kind of priority than today's.

There are two possibilities for organisational change. The first is simply to spend the money on upgrading libraries and buying national licences to more digitised information. This would simply mean upgrading the system with which we are familiar today, and will be insufficient for tomorrow's world.

The other extreme would be to establish a centralised national body of research libraries which would decide on common standards, negotiate licences and which would leave a shrinking budget and a great deal of frustration for the local branches of the national digital library. This model would kill local initiative and hinder further development.

There is a parallel to the development of digital libraries in the development of credit card organisation. Initially, each bank had its own card with a low number of transactions and a high level of operating expenses. The perspective in credit card technology was – as we know – a global card that could be used by all banks. The conditions for such a card had to include technical standards and standards for use and a code of ethics. From this platform, credit cards rapidly became global and the speed of increase and development peaked even though the card was, and is still, administered by a very small organisation.

In relation to libraries, a national digital library can be organised along very simple rules: no library is obliged to join it, but any library may as long as it accepts the rules, all participants can and should be innovative in relation to the common services, the digital library should be owned and run by the participants. The cornerstones of the model are global IT, networking models of the kind found in Silicon Valley and the Danish tradition for co-operative societies.

The crux of Denmark's Electronic Research Library project is to obtain markedly improved service for users which will, in turn, strongly encourage the libraries to continue the service and the co-operation.

THE FUTURE OF MATHEMATICAL DATABASES

Table of contents

This article was written by Jean-Pierre Bourguignon and the European Mathematical Society in the context of the work of the OECD Megascience Forum.

OECD 1999

I. INTRODUCTION

All scientists are aware that the communication, publishing and archiving of scientific work is currently going through a period of dramatic and unprecedented change, thanks to developments in electronic technology. While change is all about us, it is probably fair to say that we still cannot see where this electronic revolution will lead as far as the publication of journals and their archiving in libraries is concerned. The final steady state, which will emerge early in the 21st century, may vary enormously from one scientific discipline to another. However, while the traditional scientific journal is plainly under threat unless it adapts to the new environment, it is already clear that one of the great beneficiaries of the new technology is scientific databases. For the first time, it is now technically feasible to build comprehensive electronic databases within a given discipline, which are equipped with efficient and flexible search procedures, and which can become a basic working tool for all researchers in the field around the world. We believe that now is the moment for each scientific discipline to seize this great opportunity to create and sensibly manage a structure which will be accessible to all in the field, and which will undoubtedly deeply and subtly influence the long-term evolution of the discipline. This article addresses the complex issues underlying this challenge for the field of mathematics. The search for rigorous proof lies at the heart of all mathematical research and, as proofs are always embedded in larger mathematical structures, it is vital for the working mathematician to be able to cite previously published literature setting out the detailed verification of the validity of these structures. It is not uncommon in some areas of mathematics today to cite papers published 100 or more years ago. In this sense, mathematical databases of past publications are probably somewhat different from the databases needed in other areas of scientific research. Nevertheless, we believe that many of the general issues and principles which are set out in this report are common to most scientific disciplines.

II. GENERAL ISSUES

The revolution in scientific publishing

New electronic technology is currently bringing about a profound revolution in the communication and publication of mathematical research. This is presumably true for all the major fields of science, although this article will concentrate primarily on mathematics because it is in this area that the European Mathematical Society has expertise, and because mathematicians use their publications in a different fashion to other scientists. However, many of the general issues and principles are clearly common to most major scientific disciplines.

It is probably no exaggeration to draw parallels between the current electronic revolution and the dramatic changes in the dissemination of knowledge produced by the invention of the printing press in 15th century Europe. Some of the specific changes taking place in mathematical publishing today are listed below.

Changes in printing

In typography, traditional hot-lead typesetting has been almost totally replaced by digital typesetting, in which the printer is controlled by a computer programme. This led Donald Knuth to create the TEX programme, which enables the author of a mathematical paper to in effect do his own typesetting to a very sophisticated standard at the same time as he writes the paper. More and more mathematicians are using TEX and its subsequent variants to write their manuscripts, and it is surely only a matter of time before this becomes almost universal practice.

This has been possible thanks to the familiarity of scientists in general, and mathematicians in particular, with computers, which enables them to prepare documents themselves. There are many gains from this new way of writing papers, most notably the ability to easily make changes and to freely choose the symbols and layout of mathematical formulae. As a result, the publisher now receives either a diskette or an electronic TEX file, and can print the material after making only minimal changes to ensure conformity of style throughout a given journal.

Over the last 15 years, the change to electronic typesetting has undoubtedly led to a significant fall in the cost of the production of journals. It is only fair to say that there is almost no evidence of either commercial publishers or learned societies having passed on this saving in production costs to the purchasers of their journals. In addition, the new form of electronic production means that most journals now automatically come into existence with an electronic version. Publishers are increasingly making this electronic version available via site licences to university libraries as part of a package bought by paying the subscription to the journal. One can legitimately question whether in the not-too-distant future the production of a printed paper copy of a journal will become the exception rather than the rule.

The Internet

The creation of the Internet has revolutionised communications between individual mathematical researchers, and the dissemination of both informal and formally published manuscripts. International communication has always been high between mathematicians (witnessed, for example, by the creation of the International Mathematical Union in the late 1890s, long before most other scientific disciplines had developed international bodies), but the Internet has provided a new medium *par excellence* for mathematical researchers around the world to informally ask each other questions and exchange preliminary versions of manuscripts which may often not be destined for any permanent form of publication. It has also become increasingly common for mathematicians to make their pre-prints freely available for all to read on Web sites long before they are formally published.

Needless to say, this practice gives rise to a host of new and fundamental questions about the basic conduct of scientific research, the relevance of publishing houses and all more formal forms of publishing, and the archiving of scientific publications. Should the simple placing of an article on a Web site be regarded as some form of publication for the purpose of attributing scientific priority (the issue is rendered more complex by the fact that the author may make changes to such an article at any time)? Will it soon become the norm that formal publication of an article has nothing to do with the communication of its mathematical contents, but is really only relevant for the career of the author and for archiving of the article? Will Web publication bring about the demise of the traditional mathematical journal and its publishers, and instead lead to some *Brave New World* in which individual mathematicians end up in effect publishing their own work on the Internet?

It is impossible to foresee at present the answer to these types of question, or to predict with any certainty the full ramifications of the electronic revolution for mathematical publishing. But they illustrate clearly the depth of the changes which the publication of mathematical research is currently going through.

The role of databases

In this section, we shall discuss both why the research mathematician needs databases and why this need has become more acute today, thanks largely to the electronic revolution in publishing.

Basic principles of writing mathematics

We should stress that there has always been a need for databases in mathematical research. First, mathematics is arguably the oldest rigorous science and one of its most remarkable features, which is probably less true of most other modern scientific disciplines, is that material published many years earlier can be as relevant for a current piece of research as papers in the contemporary literature. Second, it must not be forgotten that the search for rigorous proofs lies at the heart of all mathematical research. Consequently, it has always been one of the basic tenets of writing a mathematical paper that the author must specifically identify sources in the published literature for all major results which he uses without providing a detailed proof. Third, it has always been considered essential good practice for the author of a mathematical paper to cite and take due account of all closely related earlier publications. This basic obligation remains one of the cornerstones of mathematical research today, and its erosion would do incalculable harm to a discipline which prides itself on the rigour and cohesion of the new theories it creates.

New publication trends

The historic need for databases has been increased by the explosive growth in the number of journals and all forms of publication which has grown both out of the increasing number of mathematical researchers around the world and the electronic revolution in publishing. In many ways, the cosy traditional world of mathematical research which existed up until about 1980, in which most mathematicians could keep track of publications in their field by regularly scanning the contents of a fairly stable list of journals, many of them of long pedigree, on the shelves of their university library, is now being profoundly modified. Indeed, there has been a sharp increase in the number of journals published over the last 20 years.

The reasons for this increase are more complex than the simple fact that there has been a steady growth in the number of research mathematicians. One factor has certainly been the profitability of journal publication for both commercial publishers and learned societies, due to the advent of electronic typesetting. A second reason has been the growing specialisation of scientific research, which has led to the creation of many journals devoted to rather narrow sub-fields of mathematics. A third reason has been the relentless pressure exerted on all academics to concretely demonstrate their research output in terms of publications to a whole tier of evaluating bureaucracies, ranging from their own university administrations, to research grant awarding bodies and central government funding agencies. It is inevitable that such bureaucracies end up by judging what they consider to be good research output in terms of statistical norms that are heavily biased in terms of the quantity and regularity of publication, rather than quality in any absolute sense of the word. Yet, in mathematical research, as in all areas of pure science, it is still as true as ever that it is the odd really original paper which really moves the subject forward and opens up important new areas of research.

History of mathematical databases

Before briefly describing how the working research mathematician uses an electronic database today, it seems relevant to briefly describe the history of mathematical databases.

The first mathematical database was the *Jahrbuch über die Fortschritte der Mathematik*, which was published from 1868 to 1942 and whose underlying principle was that each volume contained all mathematical research publications appearing in a given year. The material published in the *Jahrbuch* was of great value to mathematicians, and remains so as shown by the fact that an electronic version is currently being prepared (see below for more details). However, the ever increasing quantity of mathematics published annually led to delays in the production of the *Jahrbuch* and in 1931 Springer Verlag began publishing the review journal, *Zentralblatt für Mathematik und ihre Grenzgebiete*.

This was followed in 1940 by the publication, at the instigation of a number of emigrant German mathematicians, by the American Mathematical Society of the review journal *Mathematical Reviews*. The two review journals began to evolve into fully fledged electronic databases in the 1980s, and they now represent the two major mathematical databases in the world. A more detailed description is given below.

Going electronic

Today, most mathematicians already use databases in their electronic form because of the efficient search procedures they provide. Typically, the working mathematician will tend to use the database to either find or verify the precise reference of a paper he needs to consult or cite. In these circumstances, he will usually be able to make a search giving both the name of the author and the title of the paper. If there is a review or summary of the paper in the database, this can also prove very helpful in deciding whether or not it is necessary to look up the full text of the article by going to the library or by downloading its electronic version (available by a subscription to the journal or by pay-per-view). Sometimes the researcher may only know the name of the author or has a vague feeling that a given author may have written something of relevance to his project. In this case, he can make a search in the database for all the articles written by the author in question over a given period of years. Many other refinements of searching are already available. For example, one can add one of a standard list of subject classifications, or specify the type of publication as book, journal or proceedings. It seems only a matter of time before mathematicians will use databases for what amounts to genuine browsing through the literature on a given field or problem.

Criteria for a good database

It is clear that the utility of a database depends crucially on three features:

– First, the database must be as complete as possible in the sense of containing virtually all mathematical publications appearing throughout the world over a given period of years. This basic feature is certainly becoming easier to realise thanks to the fact that most journals are now produced in an electronic form. As a result, it is now technically easy for a publisher to transfer the contents of a given journal in either complete or limited form to the database (for example, the publisher could transfer the table of contents, with an author summary of each article at first, and the full text later if needed for reviews). Of course, publishers themselves are already producing smaller electronic archives covering their own publications, but it is self-evident that such limited archives can never play the full role of a truly global database.

- Second, the database must have efficient and flexible search procedures. Indeed, it is no exaggeration to compare the function of the search procedure of the database to the spinal cord of a living creature. Much work is currently being done to develop software for a better metastructure of search procedures for mathematical databases, most notably by the MathDocCell in Grenoble (see below).
- Third, the material stored in the database on a given item must be in a form which is accurate and readily useable both for search procedures and by the mathematician consulting the database. This third criterion makes direct human involvement in the classification and organisation of material in the database absolutely essential. Some discussion has occurred over the possibility of creating scientific databases simply by machines making statistical analyses of the full text of each article (the so-called "full-text searches"), but it seems inconceivable that such a scheme could work in mathematics, for example in defining the main subject classification of the paper, or in choosing certain articles for review by someone other than the author.

The ownership of databases

It should be stressed that the large electronic databases created today in a given scientific discipline will have a very long lifetime, and it may well never be easy to create new and fundamentally different ones at some later stage in the evolution of the discipline. Thus it is highly likely that these databases will end up by themselves exercising a subtle and important influence on the long-term evolution of research in the field. Therefore the question of their ownership appears to be one of the key long-term issues of principle facing all major fields of scientific research. This is especially true of mathematics, where the use of past references is a cornerstone of the discipline. Also, scientists must have genuine familiarity with the way in which their databases are built, and this will only come about if they have their say in the running of the databases. We already witness on the Internet how exactly the same query can give totally different results because of the ways in which different search engines structure their data.

It is unquestionably expensive to build and maintain large databases (see below for an analysis of the current costs of *Mathematical Reviews* and *Zentralblatt-MATH*). It is also true that commercial interests play an important role in the operation of current databases, and will continue to do so in the future. However, two principles stand out as almost self-evident for the long-term future of scientific databases.

The first principle is that the ownership of all major databases should be clearly vested in the scientific community which produced much of the material they contain in the first place. Such a principle, of course, in no way excludes the full participation of commercial partners in the day-to-day running and distribution of the database. Secondly, subscriptions to databases should be set at fair levels after a transparent analysis of costs, and the scientific community of users of the database must then be prepared to meet these costs themselves. It may well be that general public funds will be needed from time to time to make major modifications to and/or upgrade a database. Ultimately, databases will only flourish if they have the full and active participation, both scientific and financial, of the scientists who use them.

Who controls the mathematical databases ?

The specific situation in mathematics is as follows.

Mathematical Reviews is owned by the American Mathematical Society, and thus is in effect owned by the North American mathematical community.

Very recently, new arrangements have been put in place for the ownership of *Zentralblatt-MATH*. In the future, four parties will own the copyright and control *Zentralblatt-MATH*, namely Springer Verlag, FIZ Karlsruhe, the Heidelberg Academy and the European Mathematical Society. These four bodies have jointly agreed to do all they can to further develop *Zentralblatt-MATH* as an international mathematical database emanating from the European mathematical community. A first major step in this direction has been carried out in France by the creation of MathDocCell at Grenoble as a joint venture by the CNRS, the Ministry of Education and the Joseph Fourier University. This unit has made important contributions to the development of software for databases (see below). In addition, in line with its national mission within France, it has improved the flow of French publications into *Zentralblatt-MATH*, and access to *Zentralblatt-MATH* by French institutions (in particular it has incited smaller groups of mathematical researchers to form consortia for purchasing subscriptions). This is only a first step, and the goal is to have similar production units in many European countries.

Hopefully, significant new areas of co-operation between *Mathematical Reviews* and *Zentralblatt-MATH* will evolve in the future. But it is also true that mathematicians and mathematical research can only benefit from having two first-rate databases, with different traditions and history, belonging to the two most important mathematical communities in the world and run for the benefit of mathematicians everywhere.

The role of publishers and libraries

Uncertainties about the future of journal publishing

The electronic revolution in publishing has brought many benefits to publishers, but it has also provided major challenges because of the difficulty of predicting the full ramifications of this revolution. It is clear that much of the future as far as the publication of journals is concerned will be determined by electronic forces. For example, we have little idea at present as to how much longer it will be economically feasible for publishers to print and distribute a paper copy of most of their journals. On the other hand, publishing journals in purely online form poses a host of basic issues for both publishers and users. There is the obvious danger for the publisher of big reductions in the number of subscriptions for online journals, especially if university libraries in a given country or region move to pool their subscriptions. But there are equally serious drawbacks for the user. For example, does the payment of a subscription to a publisher for an online journal for a given period guarantee the user some form of permanent access to the journal even after the subscription has expired? This last issue is very important for mathematical journals because of the constant need to consult and cite past literature, as explained above.

Collaboration between publishers and databases

It seems that the best long-term answer for many of these difficulties is for publishers to accept the principle of collaborating closely with large central databases which are owned by the relevant community of scientists and therefore fall outside the normal circle of economic competition.

An ideal scenario of such a collaboration might be as follows. At the time of publication of a journal, the publisher would transfer limited data about its contents (for example, a table of contents and either an author summary or the first printed page for each article) to the central database, which would immediately make this information available to all of its subscribers as a current awareness service. This material would also be used for the normal functioning of the database, including the classification of the material and the choice of reviewers where appropriate. Such a transfer would in no way compromise the vital commercial interests of the publisher, as those who wish to see the full text of the journal would need to buy a normal subscription. It could also be argued that placing the table of contents of the journal in the current awareness section of the database would provide good publicity for the journal.

Finally, after enough time has elapsed for the publisher to be confident that subscriptions will not be lost (*e.g.* a period of several years), the publisher will transfer the full text of the journal to the central database. Once this had been done, it would become the responsibility of the database to ensure that the journal is kept electronically in perpetuity. In fact, *Zentralblatt-MATH* is planning to launch just such a scheme, called *Current Awareness Programme* (CAP), in collaboration with the European Mathematical Society, and it is hoped that it will quickly evolve to include many publishers.

Libraries and archiving

It is equally clear that the electronic revolution in publishing is going to have a profound impact on libraries.

First, the need for the archiving of electronic scientific works becomes even more acute when there is no print equivalent. Even if some publishing houses decide to maintain extensive archives of their own electronically produced journals, publishing businesses do not last eternally and will be subject to commercial forces, which will be largely concerned with the current and future activities of the publisher. Thus leaving the archiving of electronic material to the publisher rather than libraries would present a grave danger that some of this material might be lost in the long term.

Second, it seems likely that the electronic revolution will make it much easier for libraries to pool resources because of the ease of distribution of electronic data. Thus one can imagine a future in which a national library or a large regional library would purchase a subscription to a scientific journal and then distribute it to a whole range of users and institutions. Presumably this library would then act as an archive for holding the full text of the journal. All of this strongly suggests a centralisation of libraries in the future. It should also be pointed out that the catalogues of major libraries will almost certainly evolve in the direction of electronic databases, and may eventually be largely indistinguishable from a collection of overlapping databases covering all the main subjects in the catalogue. Thus, collaboration between major scientific electronic databases and libraries is very likely to become much closer, and there will be a need to carefully work out models for co-operative arrangements.

Finally, libraries will also have to cope with the complex problems of intellectual property arising from the electronic revolution, in particular, deciding on the establishment of a legal basis for publication, copyright and deposit.

III. DATABASES IN MATHEMATICS

The international mathematical community is fortunate in that it possesses two fully independent, flourishing and comprehensive databases, one centred in Europe, *Zentralblatt-MATH*, and the other in North America, *Mathematical Reviews*. It is one of the curious twists of mathematical history in the 20th century that both databases were established by the same person, Otto Neugebauer, who fled from the Nazi regime in Germany to the United States in the middle of his academic career.

Today, it is fair to say that these two databases both compete and collaborate in areas which do not affect the sovereignty of either party. The following sections give a current description of both databases and some related projects.

Zentralblatt-MATH and *Mathematical Reviews*

As mentioned earlier, *Zentralblatt fur Mathematik und ihre Grenzgebiete* was founded in 1931. It has just been renamed *Zentralblatt-MATH* to emphasise its evolution into a fully fledged electronic database. It is published by Springer Verlag and edited by the European Mathematical Society, FIZ Karlsruhe and the Heidelberger Akademie der Wissenschaften. *Mathematical Reviews* was founded in 1940 and is published and edited by the American Mathematical Society. Both journals began to develop into electronic databases in the 1980s.

Zentralblatt-MATH is based in Berlin as a separate service of FachInformationZentrum (FIZ), whose headquarters are in Karlsruhe. *Mathematical Reviews* is based in Ann Arbor, Michigan. Both databases receive all journals and books which they analyse and review at these addresses. Both databases aim to be as comprehensive as possible, covering all areas of pure and applied mathematics, as well as some mathematical material from related disciplines such as computer science, mathematical physics, economics, engineering and the biological sciences. The analysis of the raw material received by the databases is carried out by teams of scientific editors working in Berlin and Ann Arbor. The first task of these editors is to assign a standard subject classification to each item received.

Happily, since 1971 the two databases have worked together to develop a common system of classification known as the Mathematics Subject Classification (MSC). The editors decide how each item should be handled in the database, for example whether there should be an independent review, an author review or simply some type of standard summary of the article. Both databases strive to have as many independent reviews of items as possible, and it should be stressed that the reviewers come from around the world. The Annex to this article gives the geographical distribution by country of the 5 000-odd current reviewers of *Zentralblatt-MATH*. The contribution of the individual reviewers to the building of the databases is immense, and it is this feature which makes these databases distributive and of maximal utility to the working mathematician.

The final product of these labours is the publication of over 60 000 items annually by each database. These items represent vital information about articles drawn from over 2 000 journals, serials, books and conference proceedings.

Both databases are available in three publication media, namely the traditional printed version, via online access and via CD-ROM. The current price per annum of *Zentralblatt-MATH* (for two of the three media) is about DM 10 000 and the current price of *Mathematical Reviews* (for two of the three media) is about US$6 500.

Incorporation of earlier mathematical publications in electronic databases

The most important project to date for including older mathematical publications in an electronic database and the electronic archive of classical mathematical literature is the *Jahrbuch über die Fortschritte der Mathematik*. It is funded by the Deutsche Forschungsgemeinschaft (DFG), and supervised by the Technische Universität Berlin (represented by B. Wegner) and the Staats und Universitätsbibliotek (SUB) Göttingen (represented by E. Mittler), in close co-operation with *Zentralblatt-MATH*. For brevity, we shall simply call this the JFM Project.

The *Jahrbuch über die Fortschritte der Mathematik* was the first reliable and comprehensive archive of mathematical publications and the forerunner of modern databases. One volume per year was published from 1868 until 1942, and each volume attempted to review all mathematical research publications appearing in the given year. Altogether, about 200 000 items are covered in the *Jahrbuch*. The editing was carried out by the Prussian Academy of Sciences with the aid of an international board of reviewers, and it was published by Walter de Gruyter Verlag. The *Jahrbuch* was the only archival publication in mathematics until *Zentralblatt-MATH* first appeared in 1931. The absence of any archival publications before 1868 will clearly make it very difficult to include earlier publications in an electronic database.

There are many reasons why the printed form of the *Jahrbuch* is no longer satisfactory for the current needs of both mathematical researchers and historians of mathematics. Firstly, it is not easy to find complete sets of the printed *Jahrbuch*, because it ceased publication long ago. Secondly, there is an author index with every volume, but there is no global catalogue and very incomplete subject indexing. Also, there is no information on the relevance of an item to subsequent mathematical research.

The JFM Project is divided into three phases. The first step is to produce an electronic database of the existing material in the *Jahrbuch*. This database will have the same structure as the electronic version of *Zentralblatt-MATH*, and users will be able to access both services simultaneously. The editors of this database are K. Dennis (Cornell University) and B. Wegner (Technische Universität Berlin). The second step will be to enhance the use of the *Jahrbuch* database with the assistance of an international group of experts. Internal linking of the data will be established, English keywords and the standard Mathematics Subject Classification will be assigned to each item, bibliographic standardisation will be carried out, the location of the full texts of items will be indicated and, when available, comments on the value of an item for the development of mathematics will be added. It is hoped that this material will continue to be updated and enhanced as time goes by, reflecting the fact that databases must evolve like living organisms if they are to be of maximal use to current research. The third step will be to identify items reviewed in the *Jahrbuch* which turned out to be particularly important for subsequent mathematical research. In addition, it is hoped to identify material which is very difficult to find in printed form (for example, several of the journals published in the period 1868 to 1942 used paper of such low quality that their long-term durability is in doubt). It is planned to digitalise the full texts of all of this material, covering about 10% of all items reviewed by the *Jahrbuch*. This electronic archive of papers will be stored at the SUB Göttingen, and will be linked to the *Jahrbuch* database.

The JFM Project is a classic example of how public funds can be used to help build a scientific database. Commercial publishers were not interested in such a project for the basic reason that it would have resulted in a product which would almost certainly have been too expensive for widespread use by the mathematical community. Thanks to the use of public funds, it is hoped to make the JFM Project available to all mathematicians at very moderate cost both as an online service and as an offline database stored on a CD-ROM.

Software and the development of efficient search procedures

It has already been mentioned that the software used to carry out efficient and flexible search procedures, and also to incorporate new data, in an electronic database plays an analogous role to the spinal cord of a living creature. This software must be tailored to the needs of research mathematicians, and it must be prepared to evolve as mathematical tastes and habits change. Because of the need for the specifications of such software to be user driven, it should be held in the public domain, and should not be owned by commercial bodies or be under the control of any kind of monopoly.

Requirements of retrieval software

Existing software can easily handle databases of the size of *Zentralblatt-MATH* and *Mathematical Reviews*. Thus size is in itself no real issue for the software. Beyond general user friendliness, there are various other features which are of crucial importance for the software. First, the software must be able to interact with existing standard software (for example, “HTML”, “postscript” and “PDF” for Internet users, and “TEX”, “BIBTEX” and, soon, “mathml” for the writing of mathematics). Second, it must be able to run on different operating systems such as “Unix” and “Windows NT”. Third, it should be flexible enough for users to develop aspects related to their own needs and operating systems, and to feed this new software back to the core programming team. An excellent model for such interactions between users and creators is provided by the evolution of TEX over the last 20 years. Finally, it must be able to run on a wide range of computer hardware.

Continuous enlargement of the database

The software must also be able to cope with the fact that the data it works on is increasing in many different ways with the passage of time. New material from mathematical journals will be continuously added to the database. Sometimes, whole new databases will be added, as in the case of the electronic version of the *Jahrbuch* currently being prepared. There will also be needs for the software to accept new fields of search, for example to differentiate between reviews and abstracts of journal articles, and to insert the full original text or the translation of an item. It should also take into account the increasing possibility of creating hyperlinks between items within the database and with other databases, for example electronic versions of journals.

Work done by MathDocCell

The work done by the MathDocCell in Grenoble has been aimed at producing software meeting the broad requirements described above. So far, it has concentrated on developing both the search procedures and the types of queries that can be inputed into the database. The current version of the software can be used on a range of hardware and operating systems, as is clearly shown by its installation on a good number of sites around the world. The next step will be to develop its local adaptability to enable interactions to take place with local server administrators and local databases. It is hoped to finish this second phase by the end of 1998, with the new software quickly being placed in the public domain. Much work is also being done to create software that will facilitate the input of material into the database. For example, it is hoped that this software will enable publishers to automatically input the table of contents and summaries of articles in their journals into the database. It will also enable different editorial offices to work simultaneously on the contents of the database without harming its scientific coherence and consistency. Finally, it must also be adaptable to the rapidly changing practices of mathematical publication, ensuring that all forms of publication end up in the database as quickly as possible.

Annex

Reviewers for *Zentralblatt-MATH* by country

The total number of reviewers exceeds 5 000 worldwide (to deal with about 60 000 notices per year).

	Number of reviewers		Number of reviewers
Germany	1 302	Chinese Taipei	22
United States	1 227	Mexico	21
Russian Federation	353	Singapore	18
Romania	333	Cuba	15
France	313	Hong Kong, China	15
Poland	302	Turkey	15
United Kingdom	255	Latvia	14
Japan	251	Portugal	14
Italy	246	Vietnam	14
India	223	Croatia	12
China	215	Ireland	12
Canada	208	Republic of Moldova	12
Czech Republic	197	Egypt	11
Ukraine	118	Korea	11
Netherlands	116	Saudi Arabia	10
Austria	113	Nigeria	9
Spain	108	Uzbekistan	9
Hungary	105	Armenia	8
Bulgaria	101	Azerbaidjan	8
Australia	93	Kuwait	8
Slovakia	74	Kazakhstan	7
Switzerland	69	Venezuela	7
Greece	68	Algeria	5
Belgium	64	Slovenia	5
Brazil	59	Brunei Darussalam	4
Georgia	52	Chile	4
Israel	49	Islamic Republic of Iran	4
Sweden	46	Tunisia	4
Belarus	44	Cyprus	3
Former Yugoslavia	34	Iceland	3
Denmark	33	Tajikistan	3
New Zealand	31	Colombia	2
Estonia	30	Jordan	2
Norway	27	Luxembourg	2
Lithuania	26	Mongolia	2
South Africa	24	Morocco	2
Argentina	22		

As well as a number of countries with one reviewer.

INFORMATION INFRASTRUCTURE ISSUES IN THE SOCIAL SCIENCES

Table of contents

This article was written by William Sims Bainbridge, Ph.D, of the National Science Foundation (e-mail: wbainbri@nsf.gov). The views expressed in the article do not necessarily represent the views of the National Science Foundation or the United States.

OECD 1999

I. INFORMATION INFRASTRUCTURE ISSUES

The development of information infrastructure in the United States at present could be described as explosive growth and revolutionary transformation. The wide-ranging revolution taking place in communications, computing and information science is entering a new phase. The contributions from the social and behavioural sciences are becoming increasingly central to progress, and advanced technology is transforming the methodologies of the human-related disciplines themselves. In this rapidly evolving context, the very definition of infrastructure changes.

Traditionally, the word *infrastructure* has been defined in two ways. In both civil engineering and the military, *infrastructure* referred to the permanent physical installations required for major activities. For example, a city's infrastructure would include highways, sewers, the electrical power grid and the telephone system. However, in the social sciences, *infrastructure* referred more generally to the underlying framework or fundamental resources shared by a community or organisation, including non-physical as well as physical public goods. Thus, the educational infrastructure includes not only school buildings, but also textbooks, the expertise possessed by teachers, the formal organisation of educational institutions, and the shared cultural assumptions in the community that support the value of education.

When applied to information technology, the term *infrastructure* sometimes narrowly means large shared physical installations, such as the cables and routers of networks, mainframes or supercomputers. However, as computer science has taught us, software is just as important as hardware. The issue for governments is where to place their major investments in creating the facilitative computing and communications environments of the future. From that perspective, one must consider infrastructure as broadly defined, notably including hardware, software, multi-use and multi-user databases, and the social organisations that create and apply them.

The information revolution today is as much commercial as it is scientific, and privatisation of Internet underscores the fact that much of the physical information infrastructure of the future will be owned and operated by corporations rather than by governments. In the United States, several corporations are now poised to offer integrated telephone, cable television and Internet services, and the demand for bandwidth will skyrocket with high-definition television plus music and movies on demand. The distinctive contributions of government are to support the fundamental scientific research upon which the future information infrastructure will be based and to employ scientific communication as the challenging test bed on which to develop the necessary technologies.

From the standpoint of the social sciences, fast Internet communications between major universities and research facilities are highly desirable, but their chief value is moving the world towards the day when an equally good connection will be enjoyed by every primary school, place of business and home. The research laboratory of the social sciences is society at large, so data must be collected from far and wide. The student body for the social sciences consists not only of those enrolled in major research universities, but also government policy makers, the mass media and the general public. Thus the new era in the development of social science information infrastructure demands fresh thinking and ambitious research.

II. PHYSICAL INFRASTRUCTURE

The current physical infrastructure for computing-related science is substantial in the United States, but it is important to ask how rapidly the capabilities of individual components will improve. For example, the government must decide how much to invest in pushing the maximum speed of supercomputers. Industry makes extensive use of advanced computing, for example the Boeing Corporation employs entirely computational methods to design and prototype its new aircraft. However, government has been practically the sole customer for supercomputers, and further advances in this area cannot be expected without continued government investment. Scientific applications of supercomputers include modeling the formation of galaxies, analysing the complex processes that determine climate and weather, designing highly efficient engines through precise simulation of combustion and employing highly iterative techniques to determine the conceptual structure of large corpora of written text. In time, the technologies developed for supercomputing, such as massive parallelism, will diffuse to educational and industrial applications and into consumer products, but manufacturing companies might never develop such systems without government leadership.

Recently, the fastest computers in the world have moved into the so-called *teraflop* range, achieving sustained, useful speeds of more that one trillion (1 000 000 000 000) floating point arithmetic operations per second. The United States Department of Energy has taken a lead with its Accelerated Strategic Computing Initiative, but the National Aeronautics and Space Administration, the National Institutes of Health and the National Science Foundation are also co-operating in this effort. A speed of four teraflops has been approached, and there has been serious talk about achieving 40 teraflops within five years. Indeed, the objective of current research is to shift away from *teraflops* to the *petascale*, which implies not only a long-term goal of one quadrillion (1 000 000 000 000 000) floating point operations per second but also rapid-access data storage on the order of one quadrillion bytes.

Social scientists are especially interested in the social organisation required to make effective use of such machines. Recently the National Science Foundation revolutionised the organisation of its supercomputing centres to create networked partnerships which have supercomputers at their hubs but also integrate a range of hardware facilities and scientific teams at numerous distributed sites. There are now two Partnerships for Advanced Computation Infrastructure (PACIs). One, the National Computational Science Alliance, is based in the old Illinois supercomputing centre which was famous for developing the first full-featured Web browser, Mosaic, the parent of Netscape. The other, the National Partnership for Advanced Computational Infrastructure, is based in the old San Diego supercomputing centre. Each PACI is a network of about 50 universities and corporations. A scientist at one location may require software compiled at another location and run on a supercomputer at a third location to analyse data stored at a fourth. Occasionally, a computer run will be done simultaneously at several locations in parallel.

NSF's National Center for Atmospheric Research includes in its infrastructure not only a supercomputer but also the aircraft required to collect data, a good example of the diversity of large physical equipment used in scientific research. The atmospheric research of the 21st century will require not only supercomputers for modeling, but also essentially real-time collation of more than a terabyte of data daily from aircraft, ships, earth satellites, and ground stations, delivered to scientists around the world. Like the wind and rain, science does not stop at national boundaries.

Much effort is also being invested in development of faster networks. The original ARPANET used ordinary telephone lines with specially adapted computers of existing design as routers, and the Internet and the World Wide Web have largely exploited existing communication networks. However, throughout its history, significant research and development efforts have been required to create the modern Internet, and the current explosion of commercial interest has not ended the need for government involvement in the development of network technology. NSF, and the US government more generally, are committed to pushing the bandwidth and other capabilities of networks, under rubrics such as the Next Generation Internet. Furthermore, NSF and other science agencies will aggressively support fundamental research in such diverse areas as the technology of 100% optical networks and optimisation of pricing schedules. Even if industry eventually creates adequate networks for commercial purposes, there is the issue of who will pay to enable scientists to use them.

NSF has invested heavily in such net-related development efforts as the gigabit test bed, asynchronous transfer mode hardware and wireless networks. Notably, NSF entered into a partnership with the MCI corporation to build the very high performance Backbone Network Service (vBNS), which is designed to connect the two PACIs with a hundred major research institutions. Currently the vBNS is improving from 622 megabits per second to 2.4 gigabits per second. In November 1997, Singapore was officially welcomed to the vBNS, and a very high priority should be placed on providing reliable, high-bandwidth connections to all scientific facilities around the world.

Under the banner of Internet2, the University Corporation for Advanced Internet Development is working with several industrial partners – notably Qwest, Nortel and Cisco – to build the Abilene network. The goal is to connect 60 universities in 1999 increasing quickly to more than 130, at 2.4 gigabits per second with portions running at 9.6 gigabits per second. Given the great geographic size of the United States, more than 13 000 miles of fibre optical cable will be required.

Naturally, this implies co-operation among a wide variety of institutions in such diverse spheres as education, industry, the military, civilian government agencies and scientific associations. Internet was originally a spin-off from a military project of the Defense Advanced Projects Agency, assisted by the civilian National Science Foundation. Such spin-offs and co-operative ventures continue today. For example, the Naval Research Laboratory, which has significant responsibilities for pushing development of secure military data transmission networks, is working with Steven Spielberg's Shoah Visual History Foundation (VHF). The goal of the VHF is to videotape eyewitness accounts of the Holocaust and to construct a computerised archive to preserve them and make them available in analytically useful ways to scholars, scientists, students and ordinary citizens. More than 46 000 interviews have been conducted, and a multimedia hypertext system allows the user to go from particular segments of an interview to written documents, maps and segments of other interviews that cover the same topics. The Naval Research Laboratory has found this system to be an extraordinarily good test for the capabilities of its networks, because it requires flexible transmission of complex patterns of very large blocks of data.

As noted earlier, computer science should consider some kinds of software to be just as important parts of information technology infrastructure as are major shared hardware assets. In general, industry has invested far less in fundamental software research than in fundamental research intended to support hardware development. To be sure, the development of the Java language by Sun Microsystems and the recent establishment of a software research institute by Microsoft indicate that industry is interested, but government still needs to make substantial investment in fundamental software research and development as a public good.

The recent interim report of the President's Information Technology Advisory Committee says: "Software is the new physical infrastructure of the information age. It is fundamental to economic success, scientific and technical research, and national security. Software is increasingly important for commerce, for communication, for access to information and for the physical infrastructure of the country".

A very different question about physical infrastructure is how extensively can various sciences take advantage of networking distributed resources into multi-user, multi-use virtual facilities. Examples include online collaboratories, remote teleoperation of scientific instruments, and hypermedia linkage of information across separate data repositories. One can also ask what kinds of non-physical infrastructure require parallel investments in order to maximise the scientific payoff from physical infrastructure. Massive data sets, libraries of software modules and institutions for training researchers are among the possible categories. One can gain greater insight into the issues by examining how they are manifested in the social sciences.

III. SOCIAL SCIENCE INFRASTRUCTURE

The infrastructural challenges faced by the social sciences illustrate the choices that all of the sciences must make over the new few years. Infrastructure, in the form of widely shared databases and the Web-based collaboratories that sustain their creation and use, is sorely missing in these fields. Researchers have begun applying experimental methodologies to the study of large-scale networks of human interaction, including markets, election systems and social systems. The convergence of new computational power, software tools and the extension of broadband networks will now enable scaled-up Internet experiments with hundreds or even thousands of subjects, crossing international boundaries and lasting over long time periods – thus modelling real-world relationships far more realistically than was previously possible. Similarly, technology for administering surveys to very large numbers of respondents over the Internet will revolutionise survey research, if appropriate techniques can be developed to compensate for non-randomness of the samples.

One impressive example is Survey 2000, an extremely complex English-language questionnaire administered on the Web in November 1998 to more than 50 000 people around the world. The scientific focus was on geographic migration and regional culture, and contingent questions tracked the respondents' movements at seven-year intervals; assessed preferences for regional food, music and literature; collected data on Internet use and involvement in the local community; and even included questions in which the respondent was asked to judge the style and quality of a series of recorded music clips. Although the sample was not random, a number of demographic items were included to assess non-randomness and to some extent compensate for it. The social scientists who created the survey and are now analysing the rich set of data, note that the conventional large national surveys would never have been able to cover the same topics, because of cost and complexity considerations.

The data are only now being analysed, but a preliminary tabulation indicates that 40 747 residents of the United States responded, plus 5 339 Canadian residents. The survey received considerable publicity in these two nations, but 100 or more people responded from each of the following 33 nations as well: Argentina (279), Australia (1 480), Belgium (159), Brazil (203), China (236), Chile (162), Colombia (174), Denmark (100), Finland (128), France (368), Germany (417), Greece (358), Hungary (604), India (332), Ireland (203), Israel (299), Italy (430), Japan (275), Malaysia (283), Mexico (706), the Netherlands (292), New Zealand (467), Norway (146), the Philippines (122), Portugal (111), Singapore (344), South Africa (382), Spain (462), Sweden (232), Switzerland (184), Turkey (133), the United Kingdom (1 431) and Venezuela (168).

Four years ago, the NSF computing and social science directorates entered into a partnership to support development of the technology required for digital libraries, in collaboration with the Defense Advanced Projects Agency and the National Aeronautics and Space Administration. One ordinarily thinks of libraries as repositories for books and periodicals, but really any kind of digital data collection of use to scientists might qualify. One of the existing digital library projects was concerned primarily with educational science and news television programmes, for example, whereas another focused on maps, and a third archived and indexed the contents of scientific journals. One could imagine an anthropological or archaeological digital library devoted to three-dimensional representations of artefacts. In paleontology, all type specimens for all species should be represented on the Web in multi-dimensional form, from every collection, large and small. Clearly, a key role in this work will be played by non-governmental organisations such as the Getty Information Institute, which focuses on art history, and the American Association of Museums.

A new phase of the Digital Library Initiative has just been launched by the National Science Foundation, the Defense Advanced Research Projects Agency, the National Library of Medicine, the Library of Congress, the National Aeronautics and Space Administration, and the National Endowment for the Humanities. Also, at NSF, the Division of Information and Intelligent Systems and the Division of International Programs are co-operating to create a new International Digital Library Initiative that would build bi-national partnerships to support research on information systems that can operate in multiple languages, formats, media, and social and organisational contexts.

In the distributed governmental system of the United States, the federal government is by no means the only leader. The state of California has just begun an extremely ambitious programme, launching the California Digital Library on 20 January 1999. In the vision of Richard Atkinson, President of the University of California, this vast project will bring "a future when our libraries, at the press of a button, can come to us, wherever we are, whenever we wish."

In the social sciences, data infrastructure is perhaps the most important shared scientific resource. Many sets of survey data are employed extensively by numerous different researchers for secondary analysis, and a number of the major questionnaires were created primarily as data infrastructure. For decades, NSF has supported the General Social Survey, Panel Study of Income Dynamics and American National Election Study. Data from these three surveys are currently available freely to anyone over the World Wide Web. Three years ago, the NSF-supported Web site of the General Social Survey was launched by the Inter-university Consortium for Political and Social Research. Anywhere in the world, students and researchers can access and analyse the GSS data at this site, without cost. Well-organised codebooks list the 3 000 questions that have been included in the GSS since its inception in 1972, and a hyperlink goes from each item to tables of data and to abstracts of any of the 3 000 GSS-based publications that used the item. Statistical analysis adequate for most purposes can be done on line, and researchers who need to use their own software for more advanced analyses can download the data. Work now under way will add data from the most recent surveys and upgrade the search and analysis systems.

Many government agencies now offer Web-based data sources of value to social scientific research and education. For example, NSF's Division of Science Resources Studies (SRS) provides many of its reports and tables of data from its Web site. SRS personnel were instrumental in the creation of FedStats, a site providing easy access to statistical information at 70 agencies of the United States Government. NSF's programme in Digital Government is taking this effort even further, developing techniques that will improve management inside the agencies as well as encouraging them to provide data electronically to researchers and the general public.

A difficult issue facing governments today is how to maximise access by social scientists and government policy makers to large sets of data concerning human beings while preserving the anonymity of the people whom the data describe. Although raw data from major national surveys are available on the Web, the codes for the geographical locations of the respondents have been stripped off, so that the user cannot deduce the identities of individual respondents. The GSS and PSID give the geocodes to selected researchers under controlled circumstances so that they can attach other geographically based data for the purposes of scientific research. The Luxembourg Income Study allows registered researchers from its 25 member nations to submit on line requests for statistical analyses and returns results in a form that preserves the integrity of data about individuals.

There is some hope that automatic systems can be developed to allow users to carry out sophisticated analyses on line, while preventing them from doing the sequence of computations that could infer the identities of individual people or organisations included in the dataset. If this cannot be achieved, then a crucial part of the scientific infrastructure will have to be social organisations designed to preserve security. NSF has been exploring possibilities in these areas, recently in co-operation with the US Census Bureau and other agencies. For example, the Boston Research Data Center provides a secure environment where researchers have access to confidential micro-panel files, and a pair of new centres is being opened in California.

However, no matter how careful one is in setting up technological and organisational security systems, expanding access to larger numbers of researchers will inevitably increase the chances that one of them will violate the solemn agreements to preserve confidentially. A spectacular security breach involving confidential data could unjustly bring massive negative publicity down on the entire social-scientific research enterprise. A recent report to the President's Commission on Critical Infrastructure Protection argued that modern societies will become increasingly vulnerable to Internet attack, both because the Internet is becoming vital to commerce and communications and because it was never designed to resist determined, skilled attackers. At the same time, civil libertarians are concerned that human freedoms may be lost in an ill-advised attempt to tighten security. Thus prominent attention must be given to security issues as Internet-based scientific resources of all kinds are developed. Social-scientific research on the social, organisational and cultural conditions that preserve or threaten security can contribute indirectly to the value of future information infrastructure for all the sciences.

IV. INTERNATIONAL NETWORK FOR INTEGRATED SOCIAL SCIENCE

Within the next decade, a new kind of social science facility could be created, the International Network for Integrated Social Science. This would be a transdisciplinary, Internet-based collaboratory that will provide social and behavioural scientists at all institutions with the databases, software and hardware tools, and other resources to conduct world-wide research that integrates experimental, survey, geographic and economic methodologies on a much larger scale than previously possible. This facility will enable advanced research and professional education in economics, sociology, psychology, political science, geography and related fields.

One model for such a facility would distribute the main functions among eight to ten Web-based centres for collecting, archiving, retrieving and integrating social, behavioural and economic data. These centres would require cutting-edge computational and networking facilities, massive data storage and new analysis tools, and the development of innovative software and visualisation tools. Additionally, thousands of educational institutions, museums, government archives, and corporations would be linked in as remote sites, sharing a set of protocols on data transfer and cost recovery.

Existing social science data come from diverse sources – NSF-supported studies, US government surveys, international records, corporate research – and they are not readily available for the secondary analysis and meta-analysis that would achieve the maximum scientific gains. Web-based linkage and distribution of these databases requires that many difficult problems must be solved regarding the indexing, archiving and searching of data. It will be necessary to establish networks of researchers located in various centres to solve problems relating to data reliability, confidentiality, storage and retrieval. Also, the problems of different data formats and the need for new analytical techniques geared specifically to Web-based data must be considered.

Many challenging social-scientific questions involve interactions among much larger numbers of people than can be accommodated in conventional labs, including markets, election systems and social networks. Through a net-based distributed laboratory, experiments can be scaled up to include thousands of subjects at hundreds of remote locations. This "netlab" facility will cross regional and national boundaries, bringing new population samples into the laboratory and permitting experimental research on the impact of culture and geography. The duration of experiments will be greatly expanded to cover evolutionary processes never before studied. The educational value of conventional social-science laboratories is limited by the fact that few universities have them, so for the first time laboratory experimentation can now become part of the routine education of undergraduates in the social and behavioural sciences.

If an efficient, comprehensive Web-based archive for social and behavioural science data existed, then new research could be solidly based on earlier studies, and fresh data could be linked to the existing databases for comparison and calibration. Internet-based experiments in the social and behavioural sciences will be merged with questionnaire surveys, and both can be linked via geographic analysis to other sources of data, including census information, economic statistics, and data from other experiments and surveys. Technology for administering questionnaires to very large numbers of respondents over the Internet will revolutionise survey research. Current national surveys are so costly that the number of respondents is typically too low or clustered in too few locations to permit integration with other sources of data by means of geographic information systems. Only by moving to Web-based administration of experiments and surveys can social and behavioural scientists achieve sufficient numbers of cases to integrate across disciplines and analytical methods.

The International Network for Integrated Social Science will take the lead in developing new computational and statistical techniques for combining conventional research methodologies and transcending their traditional limitations. Many of the studies conducted through it will challenge the next generation Internet, because they require very rapid and reliable communication linking numerous sites, and the international scope of the project challenges current organisations and research paradigms. Existing methods for searching and distributing data will be challenged by the very large number of individuals and organisations that will access the databases, in science, education, government and commerce. This facility will undertake calibration surveys employing standard methods to compensate for non-randomness of samples obtained in the Internet-based questionnaires and experiments. It will create software to integrate experiments into questionnaires and incorporate both into geographic information systems. Like a great telescope, this facility will provide the physical infrastructure and organisational framework for research by scientists from many universities. Its distributed archive will provide data and statistical analysis tools to researchers and students everywhere.

In the Information Society of the 21st century, the data contained in this international network of archives will be of great value to education, commerce and the general enlightenment of citizenry. Therefore, technical and institutional means must be developed to provide the data to as wide a user community as possible, consistent with the privacy and intellectual property rights concerns that may apply to some sub-sets of data. Scientific discoveries are a public good for the entire world, so a very high priority should be placed on the availability of government data for scientific analysis. Each major archive in the network should have procedures that give accredited scientists access to otherwise confidential data, either under strict legal procedures that protect against misuse, or through technical systems that limit data output to general findings that do not compromise confidentiality.

New data collection efforts will achieve higher levels of quality by drawing upon the general findings, particular measures and practical research experience of the previous studies that have been archived. Longitudinal studies will construct time-series comparisons across data sets to chart social and economic trends. Each new study will be designed so that the data automatically and instantly become part of the archives, and scientific publications will be linked to the data sets on which they are based, so that the network becomes a universal knowledge system.

V. LEVELS OF INVESTMENT

It is very difficult to estimate the amount invested each year in the United States in information infrastructure, both because one does not start with a clear definition of the phenomenon and because investments are made by a very large number of organisations, public and private. However, clear pictures of some of the pieces can be offered to provide a sense of their absolute and relative magnitudes.

Each year, Federal agencies report how much they invested in research and development in the area of computing, information and communications, and in fiscal year 1998 the total was about US$1 070 000 000. Two agencies accounted for more than half of this investment: the Defense Advanced Projects Agency (US$ 321 million) and the National Science Foundation (US$ 284 million). DARPA and NSF were, of course, the agencies responsible for the creation of Internet. Ten other agencies accounted for the remainder: the Department of Energy (US$ 129 million), the National Aeronautics and Space Administration (US$ 128 million), the National Institutes of Health (US$ 92 million), the National Security Agency (US$ 36 million), the National Institute of Standards and Technology (US$ 27 million), the Department of Veterans' Affairs (US$ 22 million), the Department of Education (US$ 12 million), the National Oceanic and Atmospheric Administration (US$ 8 million), the Agency for Health Care Policy and Research (US$ 6 million), and the Environmental Protection Agency (US$ 5 million).

These significant sums exclude research and development in areas of communications and information management which do not involve advanced computing, and they cover very little of the actual deployment costs of operational computer networks, hardware and software. Conceptually, the US$1 070 000 000 is distributed across five categories: high-end computing and computation (US$ 462 million), (US$ 33 million), human-centered systems (US$ 280 million), and education, training and human resources in the computer science area (US$ 39 million).

The vBNS is budgeted at US$ 50 million, and the universities involved in Internet2 are investing US$ 50 million per year in Abilene and related work. Similarly, the current phase of the federal Digital Library Initiative will invest over US$ 50 million. Investment in the California Digital Library will be on the order of US$ 20 million in new state money each year, matched by money from participating universities which may be diverted from more traditional expenditures. If, as many expect, scientific journal publishing becomes all electronic within a few years, a large fraction of library holdings will gradually become digital.

Traditionally, investments in the social sciences have been more limited and a significant increase would be required to reach the goals suggested in this paper. The three main NSF-funded surveys will cost only about US$ 17 million over the four-year life of their current grants: the Panel Study of Income Dynamics (US$9 392 050), the National Election Studies (US$3 954 259), and the General Social Survey (US$3 835 000). The best current cost estimate for research, development and demonstration projects for the American part of the International Network for Integrated Social Science is much higher, at approximately US$ 100 million.

All of these numbers pale before the value of the new information infrastructure to science and to society more generally. Already by 1996, fully 64 % of American pre-college schools had some access to Internet, and half of all companies had Web sites. Internet has become the arena for vast business activities. For example, Netscape was recently purchased by AOL for US$4 200 000 000, and the Excite corporation, which began as a mere search engine for the Web, sold for US$6 700 000 000. Despite the economic weakness experienced in many nations at the present time, there is good reason to accept the rosiest predictions of the value of the world's future information system. Involvement of the social and behavioural sciences in the creation of this challenging computing and communication system will maximise the benefit for society and simultaneously enable discovery in the human-related sciences.

VI. CONCLUSION

Science is an international partnership, so one must think globally when considering infrastructure issues. Many nations manufacture computer and communications hardware, and their scientists should rise above nationalistic competition to create the physical infrastructure needed by the science of the 21st century. To the extent possible, software should be modular and compatible, so international co-operation in creation of scientific software is essential. Nations should commit themselves to making government data readily available to scientists and should enter into international collaborations for the collection of fresh data. The technology of Internet and the World Wide Web will enable new kinds of virtual research centres and global data sharing that will greatly enhance the work of scientists and educators everywhere. Progress in developing the scientific information infrastructure for the future can best be achieved through the broadest possible international co-operation and through recognising that this infrastructure combines hardware, software, databases and social organisations.

OECD PUBLICATIONS, 2, rue André-Pascal, 75775 PARIS CEDEX 16
PRINTED IN FRANCE
(90 1999 24 1 P) ISBN 92-64-16178-3 – No. 50783 1999